THE UNIVERSITY *of* EDINBURGH

Linkage and Association Mapping for Quantitative Phenotypes in Isolated Populations

Christopher S Franklin

Thesis submitted for the degree of Doctor of Philosophy

University of Edinburgh

2010

Declaration

I declare that this thesis was composed by myself and that the work contained herein is my own.

The work has not been submitted for any other degree or professional qualification.

Christopher Steven Franklin

Abstract

Many complex diseases are known to have a substantial genetically heritable component. Elucidation of these genetic risk factors provides increased knowledge of the biological mechanisms that result in the diseases while also presenting new potential targets for therapy. This thesis explores the methodology of mapping genetic loci using isolated populations in the context of quantitative trait analysis.

Chapter 1 explores the rational for the project, discussing the benefits of using quantitative traits rather than binary disease status and the pros and cons of using isolated populations. This is followed by a brief history of genetic mapping with reference to type 2 diabetes mellitus (T2D) and related quantitative traits.

Chapter 2 introduces the methods used in this thesis. This includes strategies to deal with medication, methods to determine kinship between individuals, linkage analysis, association analysis and meta-analysis of multiple studies.

Chapter 3 presents linkage analysis of T2D related traits carried out in 2 – 4 populations depending on availability of the traits and appropriate marker data.

Chapter 4 presents the results of association analysis for T2D related traits in 3 – 5 populations using genome-wide SNP data. The results using the alternate methods described in chapter 2 are compared using fasting glucose as this was the most widely measured phenotype.

Chapter 5 introduces additional traits derived by pulse wave analysis and discusses their relevance to metabolic disease before presenting association analysis using the preferred method from chapter 4.

An overall discussion of the strengths and weaknesses of the analysis is given in chapter 6.

Acknowledgements

The work presented in this thesis was carried out over a period of three years while studying as a postgraduate student at the University of Edinburgh. Throughout this time I was funded by an Economic and Social Research Council (ESRC) postgraduate research studentship. I would like to thank the ESRC for supporting me during this period of research.

I carried out my studies primarily in the Centre for Population Health Sciences at the University of Edinburgh and I would like to thank many people who assisted me during the period of my PhD.

I would like to thank my supervisors, firstly Dr Jim Wilson for providing both the dataset on which the majority of this thesis was based and a seemingly limitless supply of enthusiasm and optimism at times when it was needed most. I would also like to thank my secondary supervisors, Dr Sarah Wild and Dr Sara Knott, for educating me in the fields of medicine and statistics, which this thesis attempted to straddle.

A great deal of thanks is also owed to several people whose support of myself and the department in which I studied was vital to the completion of this project. Rosa Bisset is the cornerstone on which the department rests and is as responsible, metaphorically, for holding everything together as the scaffolding, which more literally held the building together. I would also like to thank Sarah McAllister and Maggie Luttrell who were consecutively responsible for the administration of my and my fellow PhD student's studies.

I would like to give special thanks to my fellow PhD students; Ruth McQuillan, Mirna Kirin and Jamie Floyd for the many discussions, scientific or otherwise, which have kept me entertained and sane for the past few years.

During some of the writing of this thesis I have been working in the Wellcome trust Sanger institute's human genetics department and I would also like to acknowledge the advice support and understanding my colleagues and co-workers there have shown me during this process.

Lastly I would like to thank my parents for their encouragement, for always taking an interest in my work, and for their unconditional support throughout this and every other time of my life.

Publication

The following publication has resulted as a direct outcome of the work presented in this thesis and is attached as Appendix I:

The TCF7L2 diabetes risk variant is associated with HbA_1(C) levels: a genome-wide association meta-analysis.

Franklin CS, Aulchenko YS, Huffman JE, Vitart V, Hayward C, Polašek O, Knott S, Zgaga L, Zemunik T, Rudan I, Campbell H, Wright AF, Wild SH, Wilson JF.

Ann Hum Genet. 2010 Nov;**74**(6):471-8

1 Chapter 1: Introduction to genetic mapping in complex disease

Wars, violence, accidents and suicide account for less than 10% of worldwide deaths each year. The remaining 90% of deaths are the result of some form of disease(WHO 2008). These diseases can be broadly divided into two categories: Those that are communicable and those that are not.

In developing regions the communicable diseases represent by far the largest health burden with, for example; according to the world health organization (WHO) 2004 report(WHO 2008), 68% of all deaths in Africa attributed to communicable disease and only 25% resulting from non-communicable conditions.

In regions with the most developed infrastructure the communicable group of diseases account for a considerably smaller proportion of annual deaths. In Europe only 6% of deaths are attributed to communicable disease and 86% to non-communicable disease.

Logically, regions with low rates of infectious disease deaths will have a higher proportion and total number of non-communicable disease deaths since, ultimately, everyone has to die of something. However, the higher proportion of non-communicable disease deaths in the developed world also represents a genuine increase in the burden caused by certain diseases compared to less developed

countries. Looking specifically at men aged 45-59; the CVD mortality rate among Africans is estimated to be 38 per 10,000, compared to 44 per 10,000 among European men; illustrating an increased rate of heart disease in relatively young men in the developed world.

From the perspective of a geneticist; even in a hypothetical controlled environment free of infectious agents or damaging environmental exposures there would still be variation in the diseases people die from and in the age at which they develop them. This is because the many common non-communicable diseases have a genetically heritable component and heritability estimates for categories such as "major psychoses, early-onset ischaemic heart disease, rheumatoid arthritis, ankylosing spondylitis and diabetes mellitus" were reported more than 40 years ago (Carter 1969). Identification of the heritable component of complex diseases is the main topic of this thesis.

The heritability of a disease or phenotype is an estimate of that proportion of the phenotypic variation that is due to inherited genetic factors and is sometimes used as a rationale for describing the importance of studying the potential genetic effects of a particular disease. It is also used retrospectively as a guide to how successful studies have been in identifying meaningful information about the role of genetic factors(Maher 2008). Heritability estimates can be calculated as either narrow sense estimates, which represent the proportion of variance attributable to additive genetic

effects, or broad sense heritability, which includes both additive effects and dominance effects.

It is important to interpret the heritability estimate in the correct context. Firstly bearing in mind that the estimate is a proportion of total variance and is therefore inversely proportional to other forms of phenotypic variance. This means that differences in levels of environmental variation will result in different heritability estimates for the same phenotype and so an estimate is only truly representative of the population in which it was calculated.

Heritability estimates are also sensitive to measurement error as this will both increase the total phenotypic variance and confound the observed covariance between genotype and phenotype(Macgregor, Cornes et al. 2006). This can result in phenotypes with relatively poor repeatability such as blood pressure or spirometric measures appearing to have less significant genetic components than more easily measureable traits but does not mean that there are not significant genetic components affecting the phenotype.

There are 6 currently known major types of genetic variation that can account for the heritable variation in a phenotype or disease. Firstly and perhaps most obviously are changes in the coding sequence of a gene and the resultant alterations of its product. A single base substitution can occur at any point in the genome and does so in each generation. The rate at which these mutations occur in the human germ-line has been

estimated in a variety of ways as far back as 1935 (Haldane 2004) but the most recent estimates suggest an autosomal base substitution rate of approximately 1.3x10-8 per site per generation corresponding to roughly 50-100 novel mutations in each individual (Lynch 2010). While germline mutations can result in phenotypic variation they will not initially contribute to the estimated heritability of a phenotype since they are only present in a single individual. They are however the ancestral source of the polymorphic sites which result in heritable genetic variance in many complex diseases.

The vast majority of these polymorphic loci will occur in regions of non-transcribed DNA since these regions make up roughly 98% of the human genome(Elgar and Vavouri 2008). However, a small proportion of polymorphisms occur within the transcribed sequence of genes.

The impact of polymorphisms within transcribed sequence is site specific and highly variable. Firstly the structure of amino acid codons, with 3 bases each of which can be one of 4 nucleotides, results in a great deal of redundancy with 64 alternative codons translating into only 20 different amino acids and 1 stop signal. Each codon is made up of 3 bases each of which can be mutated to one of 3 alternative bases so a single base substitution results in a change to 1 of 9 alternate codons. If there were an even distribution and random assignment of codons across amino acids this redundancy would give an average of 3.05 codons per amino acid and the probability of this resulting in an amino acid substitution or creation of a premature stop codon

would be 1-(3.05/64), approximately 95.2%. However, the relationship between codons and amino acids is not random or evenly distributed.

The actual number of codons per amino acid ranges between 1 and 6 with a median of 2 and groups of codons that code for a particular amino acid generally share a similar structure with the first base being most strongly conserved and in many cases the third base being completely interchangeable.

This structured coding results in different probabilities of non-synonymous changes for different amino acids. Tryptophan and methionine, which shares its genetic code with the transcriptional start site, have unique codons and so are susceptible to any base substitution. Conversely, arginine, serine and leucine each have 6 alternate codons and so the probability of a synonymous base substitution is highest in the codons for these residues.

Besides altering the protein coding sequence of a gene a single base substitution can also disrupt splice sites that will result in alteration of the splicing pattern of exons in the messenger RNA. This will result in entire sections of the protein product being absent rather than an alteration of a single residue.

Similarly to splice site disruptions, single base insertions or deletions within coding sequence can also have a large impact on the protein product produced. This causes a

frame shift of the entire coding sequence downstream of the mutation and results in the majority of the amino acid residues being miscoded.

All of the types of coding variation described can be exemplified with reference to the recessive monogenic disorder cystic fibrosis (CF). CF is caused by defects in the production or function of the cystic fibrosis trans-membrane conductance regulator, the gene encoding this protein, *CFTR,* is known to have over 1,000 functional variants with varying frequencies across populations (Rodrigues, Gabetta et al. 2008).

The most commonly observed disease-causing variant is ΔF508 (Kerem, Rommens et al. 1989), A 3 base deletion straddling the 507th and 508th codons of the *CFTR* sequence. Codon 507 changes from ATC to ATT both of which code for isoleucine, codon 508 is effectively missing resulting in the loss of a phenylalanine residue. As the deletion removes 3 bases it does not change the reading frame of the gene and so the remaining coding sequence is transcribed and translated as normal. The loss of the phenylalanine residue disrupts the normal folding of the protein produced resulting in its degradation before reaching the surface of the cell (Rowe, Miller et al. 2005).

Another considerably more rare variant in *CFTR* within the 533rd codon consists of a single base C->T substitution, which results in a premature stop codon replacing the wild-type Arginine. This causes a truncated non-functional protein to be produced.

Yet another single base substitution demonstrates the potential effect of splice site variation. A G->C transversion at the 3rd position of the 1291st codon in the 20th exon of CFTR results in a Histidine residue in place of the more common Glutamine residue. While this does represent a protein coding change it also results in disruption of a splice site (Jones, McIntosh et al. 1992). The disruption results in an additional 29 bases of intronic sequence being included in the mRNA product, and this additional sequence contains an in-frame stop codon, which causes premature termination of transcription.

The second broad category is changes that affect regulatory elements and modify the expression of one or more genes. Promoter regions are usually located near to and upstream of the gene they are associated with making them relatively easy to identify. Enhancing elements on the other hand can be located a great distance from the gene or genes they control. Disruption of either of these types of elements can lead to a change in the expression levels of a particular gene so that although the gene product is the same it may be present in too large or too small a quantity to perform its required role. The expression patterns of certain genes during development can be highly specific and so mutations that disrupt the expression control can cause strong phenotypic effects.

A third source of genetic variation comes from the fact that individual genes do not act alone but rather fall into networks and pathways. This leads to the concept of epistasis whereby variation in one gene's function or expression levels can enhance

or negate the effect of variation in another gene. An example of this phenomenon is the Bombay phenotype (Hakim, Vyas et al. 1961). This is a rare recessive condition in which an individual produces no H-antigen, which is the precursor of A and B blood group antigens. Since the person can produce no blood-group antigens they appear to be blood group O even if they inherited A or B alleles from their parents.

The fourth source of heritability comes from the interaction of genetic variants and environmental variation. In this case the potential phenotypic variation caused by a genetic polymorphism may only be observable in certain environmental conditions. For example several risk loci for alcohol dependence have been identified and confirmed in the last 5 years but it is clear that the effect of these genes will only be observable in an environment where alcohol is available and the effects may be modified by the attitude towards drinking alcohol prevalent within a given society.

The fifth category of genetic variation, which may contribute to the heritability of phenotypic variation, is Copy Number Variants (CNVs). A CNV is essentially classified as a region of the genome, which may contain one or more genes, being present in variable numbers within a population. Generally each individual has 2 copies of any section of autosomal DNA but faults can occur during replication resulting in the deletion or duplication of a particular region leading to a variable dosage effect of the genes present in this region.

The sixth and least well-characterized source of variation is heritable methylation. Methylation of regions of the genome prevents the binding of transcription factors and polymerases thus preventing expression of the genes. Although this does not result in any permanent change of the DNA sequence there is evidence to suggest that some types of epigenetic modification are passed on from parents to offspring and that this could therefore contribute to their phenotypic correlation.

Certain epigenetic changes relate to the imprinting of specific loci during germ-cell development and result in differential expression levels of alleles based on whether they were maternally or paternally inherited. Many of these imprinted loci have been identified both in animal models and in the human genome (Hirasawa and Feil 2010) and appropriately designed studies have already begun to identify potential phenotypic consequences of this imprinting (Kong, Steinthorsdottir et al. 2009). Even though this type of imprinting is reset between generations it has implications for the study of genetics by obfuscating the relationship between genotype and phenotype at these loci.

In addition to single generation sex-specific imprinting there is also evidence for histone modifications and methylation patterns at some loci that can be passed on unchanged from one generation to the next (Hammoud, Nix et al. 2009). This represents an additional source of heritable variation, with impacts on phenotypic variance, which is not yet detectable in DNA sequence analysis.

It has also been demonstrated in animal models that patterns of epigenetic variation can be inherited through the germ-line through the transmission of small RNA molecules. This type of inheritance has been demonstrated to be experimentally possible through the introduction of small RNA molecules in mice gametes that cause persistent alterations in coat colour (Rassoulzadegan, Grandjean et al. 2006) or in cardiac development (Wagner, Wagner et al. 2008) and the presence of RNA molecules in unmodified mammalian gametes means that this is an entirely plausible hypothesis for trans-generational epigenetic inheritance (Krawetz 2005).

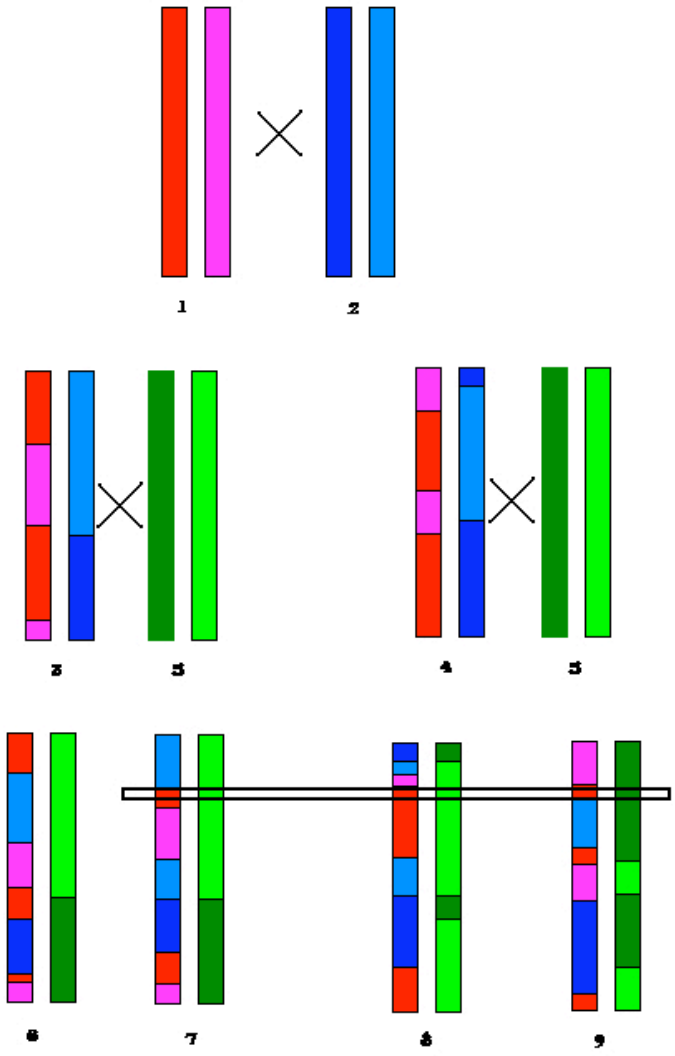
Having established that many diseases have a heritable genetic component the challenge becomes identifying more specifically which genes or genomic elements are involved in the development of the disease. The characterization of polymorphic genetic markers was crucial to this endeavor. The first markers to be used in genetic mapping studies were multi-allelic micro-satellite markers, which consist of variable repeats of the same short DNA sequence. Repeat sequences such as these have a relatively high probability of mutation between generations through slippage during replication. This leads to many separate alleles being present within a population giving high levels of heterozygosity and allowing the tracking of segments of DNA between parents and offspring.

Micro-satellites have been used primarily in linkage mapping for diseases and quantitative traits. This technique requires a population of related individuals who are measured for the trait in question. Using a genome-wide collection of micro-

satellite markers and the known relationships between individuals it is possible to draw up a map of shared genomic regions. At each point along the linkage map a probability that two individuals share the same genotype can be calculated and this can then be compared with incidence of the disease or with covariance of a quantitative trait. This highlights regions that are likely or unlikely to contain a locus with an impact on the phenotype.

Figure 1 demonstrates the simplest possible case of linkage analysis. Individual 1 is a female with a genetically heritable disease. She mates with an unaffected male "2" producing two daughters "3" and "4" each of which exhibits the same disease. The daughters then have children with unaffected men, both represented by individual "5". These unions produce a total of four children "6" who shows no signs of disease and "7", "8" and "9" all of whom have the disease. If we are able to genotype all three generations of the family then we can track the inheritance of sections of individual "1"s chromosomes through the family and see which sections are shared between the affected individuals. In this case the only section shared between "1", "3", "4", "7", "8" and "9" is a short region coloured in red and circled in the last generation. This pedigree would suggest the genetic variant underlying the disease is located in this region. Separate analysis of other families in which this disease is found may give evidence in support or contradiction of this region and a large study of several affected families would be needed to confidently identify a linked region.

Figure 1. Mendelian inheritance and linkage mapping in a simple pedigree



Linkage mapping has proven successful in animal research since the artificial selection process imposed on livestock, using a small number of males to sire millions of offspring, greatly reduces the effective population size and as a result reduces the total genetic heterogeneity with many polymorphic sites rapidly becoming fixed in the population. The ability to design specific breeding experiments also maximizes the power for this type of analysis.

The linkage mapping technique was also used with success to identify the loci responsible for some rare monogenic diseases including the first genetically defined for of maturity onset diabetes of the young (MODY)(Bell, Xiang et al. 1991) and a form of familial hypercholesterolemia (Eden, Naoumova et al. 2001). When applied to the study of complex polygenic diseases and traits linkage has been considerably less successful. The strength of linkage analysis lies in identifying loci where rare alleles account for a large proportion of phenotypic variance. However, linkage mapping has very low power to identify common variants that individually have only a small impact on the phenotype.

Association analysis works on a different principle to linkage analysis. Over the course of one generation large segments of a chromosome may be passed on intact from parent to offspring however over successive generations recombination will reduce the size of the section of DNA which is identical to the original ancestral segment of DNA. Over dozens of generations the regions of shared DNA are broken down to small blocks of a few hundred kilobases, which are still identical to the

original DNA. Loci within these sections of DNA are described as being in linkage disequilibrium (LD) within a population meaning that the genotypes at these loci are correlated with one another. The level of LD between two loci can be calculated by examining how often the same alleles at the two loci occur together. If there is high LD between two loci then genotyping one can be used to predict the genotype at the other and if LD is complete then the section of DNA in between the two loci can be assumed to be identical as well.

This process can be used to identify regions of DNA that are expected to be shared identical-by-decent (IBD) between two individuals who are apparently unrelated but who ultimately share a distant common ancestor.

Association studies generally use bi-allelic Single Nucleotide Polymorphisms (SNPs) rather than variable length micro-satellite markers since SNPs are far more numerous, have a simpler mutation process, and are much less prone to further mutation. As previously mentioned the rate of single base substitution is approximately $1.3x10^{-8}$ whereas the mutation rate for microsatellites is estimated to be between 10^{-6} and 10^{-2} (Eckert and Hile 2009).

Initial association studies used markers in and around genes for which there was some hypothesis for involvement in a particular disease. These candidate gene studies produced some positive results for example demonstrating the link between obesity and the melanocortin receptors 4 and 5 (Chagnon, Chen et al. 1997).

However, the overall results produced by these types of studies were less dramatic than many had hoped.

The two major flaws with candidate gene studies were firstly; the underestimation of how many loci would contribute to most complex phenotypes, and therefore the overestimation of expected effect size per locus, and secondly; the assumption that our prior knowledge of functional genetics would prove a strong enough guide to identifying appropriate candidates.

Genome-Wide Association studies (GWAS) attempt to overcome these two flaws. The issue of effect size is ameliorated to a certain extent because the heavy burden of multiple testing in a GWAS already requires larger sample sizes to distinguish between true results and false positives. The GWAS approach also addresses the issue of the usefulness of prior functional knowledge by attempting to test all regions of the genome. In practice the coverage achieved depends on the density of markers used and certain high recombination regions are difficult to tag using LD.

1.1 Relating quantitative traits to disease outcomes

1.1.1 Study of genetic components of common diseases

Classical epidemiology involves the discovery and investigation of risk factors involved in diseases. This type of research has determined the negative impact which modifiable risk factors such as smoking, obesity and stress have on people's health

through the development of complex diseases like Coronary Heart Disease (CHD), various forms of cancer and Type 2 Diabetes (T2D). Public health research is in turn involved in attempts to reduce the prevalence of these risk factors in the general population. Interventions which reduce the amount of smokers or which encourage exercise can be expected to improve the general health of the population in the long term.

This method is however limited to modifiable risk factors. Most common complex disorders are known to have a sizeable genetic component meaning that one individual may be born with a higher genetic predisposition to a particular condition than another. The study of the genetics of these diseases is therefore important for the purposes of understanding more fully the biological pathways that lead to disease. This knowledge can be used in turn to guide the development of new pharmacological treatments for the disorder.

Historically many drug treatments have been discovered by accident or trial and error and are prescribed in standard doses to all patients with a certain set of symptoms. Detailed understanding of the underlying biology of diseases allows new products to be purposefully designed to produce a specific biological effect. Genetic knowledge also allow for differential diagnosis of patients who may be presenting with the same set of symptoms but with different underlying causes

1.1.2 Study of disease outcomes as binary variables

GWAS analysis has been used on a variety of common complex diseases using a case-control design. The major benefit of this type of study is that large cohorts of cases already exist for many conditions and, pending ethical approval, can be used along with a suitable set of controls. It is also possible to create and genotype a set of control samples that can be re-used for many different diseases thus reducing the overall cost of genotyping as was the case with the Wellcome trust case control consortium (WTCCC 2007).

The case control design does have certain drawbacks for use in GWAS analysis. Firstly the ascertainment of appropriate controls can be difficult depending on the disease in question. With late onset conditions a healthy control may in fact be a case that has yet to present symptoms.

One strategy to overcome this is to simply use a very large set of un-phenotyped controls in which case the increased power from sample size can compensate for the effect of cryptic cases being present in the control sample. The benefit of using this strategy depends directly on the population prevalence of the study disease since this is the proportion of cryptic cases you expect to be present in your un-typed controls. It may therefore be highly successful for relatively rare conditions such as type I diabetes but will be less effective for very common conditions such as obesity or hypertension where some level of control phenotyping may be necessary.

A second drawback of the case control design is disease heterogeneity. Disease definitions are largely compiled from a set of symptoms. This is clearly useful when determining an appropriate course of treatment to address those symptoms but from a developmental perspective the same disease outcome may arise from entirely separate mechanisms with separate underlying genetic influences.

In the case of Myocardial Infarction (MI) for example, prolonged periods of high blood pressure can result in arteriosclerosis increasing the probability of an MI event. At the same time, increased blood lipid levels can result in the formation of atheromatous plaques, and increase the probability of MI, in individuals with normal blood pressure. It is therefore possible that a case control GWAS for MI could be testing for association with either blood pressure or blood lipid levels but not doing either in an efficient way.

The metabolic syndrome is another example of a disorder diagnosed using a variety of component phenotypes. Several different definitions of the syndrome have been published since 1988. The definition used by the World Health Organization (WHO) requires 3 or more of the following factors: fasting glucose (FG) $\geq$ 6.05 mmol/l, triglycerides $\geq$ 1.65 mmol/l, high density lipoprotein (HDL) < 1.04/1.29mmol/l for men/women respectively, blood pressure $\geq$ 130/85 mmHg and central obesity measured by Waist to Hip Ratio (WHR) or Body Mass Index (BMI). A case control GWAS of metabolic syndrome will in reality be testing for all of these phenotypes

but none of them are individually sufficient for diagnosis and combinations of the components are likely to differ between populations.

An alternative study design involves using continuous quantitative phenotypes instead of discreet disease categories. This technique is potentially more powerful than, and certainly complementary to, the case control study design in a variety of ways.

Firstly, the component phenotypes that underlie a disease are inherently less complex than the disease since any variation affecting the phenotype will also result in variation in disease risk. This reduction in the number of sources of variation makes identification of those sources - in this case genetic polymorphisms - easier.

Secondly the continuous phenotype may have a great deal of variation within the disease and healthy categories and this variation may have predictive value for future disease. For example, fasting glucose levels are both a diagnostic measure for T2D and, at non-pathogenic levels, are a strong predictor of future T2D risk.

In more general terms studying continuous phenotypes has economic advantages since each person has a value for any given phenotype and so the same set of people can be measured for a large number of different phenotypes. This results in a single set of genotyping contributing to studies of a cornucopia of different diseases and disorders.

For these reasons studying the continuous phenotypes that contribute to disease development can be a complementary approach to case control analysis.

1.2 Benefits of isolated populations

Isolated populations are small populations in which geographical or social factors have caused a period of genetic isolation, with little or no immigration from other populations. The use of isolated populations in the study of human genetics first became popular with respect to rare Mendelian recessive disorders. The small population sizes result in a background level of inbreeding which increases the probability of a rare allele being found in a homozygous state(McQuillan, Leutenegger et al. 2008). This is beneficial for genetic mapping strategies both in increasing the effect size of additive loci found in a homozygous state and in increasing the ability to detect dominant effects.

Historical population bottlenecks result in longer regions of linkage disequilibrium between the disease gene and surrounding markers(Latini, Sole et al. 2004). This is also beneficial for genome-wide association mapping techniques as they rely on tagging unknown causative variants with a panel of markers. The coverage achieved with a given marker panel will be larger in populations with larger LD blocks than in populations with small LD blocks(Peltonen, Palotie et al. 2000; Heutink and Oostra 2002).

In terms of linkage analysis isolated populations are also beneficial because they lend themselves to the collection of large family groups as participants encourage their relatives to join. In this way a significant proportion of the total population can be recruited and large family pedigrees can be constructed for use in linkage mapping. It has been shown that for variance component linkage analysis of quantitative traits extended pedigree study designs achieve greater power to detect quantitative trait loci (QTLs) than sib-pair studies with the same total sample size (Williams and Blangero 1999).

A theoretical benefit of using a small isolated population in the study of multi-factorial diseases and traits is that the environmental variance present should be lower than when sampling from a large metropolitan or national population(Heutink and Oostra 2002). This reduction in the environmental variance should result in an increase in the relative proportion of phenotypic variance that results from genetic variation.

The small population size and historical bottlenecks also increase the effects of selection pressures and random drift both of which can drive polymorphic sites to fixation(Pardo, MacKay et al. 2005). In the case of studying a trait that may be influenced by tens or hundreds of genes this should result in the fixation of some of these genes in a non-polymorphic state. This would then increase the relative contribution of the genes that are still variable i.e. giving them large effect sizes and making them easier to identify.

1.3 Statement of aims

The main aims of this thesis are two-fold. Firstly, to investigate methods appropriate for the analysis of high-density genetic data in isolated populations. Methods for conducting linkage analysis in extended pedigrees are well established(Almasy and Blangero 1998; Abecasis, Cherny et al. 2002). However, with the advent of high density SNP data, efficient methods for conducting large-scale association studies in pedigree data are relatively new (Aulchenko, de Koning et al. 2007; Chen and Abecasis 2007).

The second linked aim of this thesis is to identify genetic loci influencing medically relevant phenotypes within our isolated population studies, thus providing a proof of principle for the methods used.

2 Chapter 2: Methods

2.1 Methods to determine kinship

An important aspect of both linkage and association analysis studies using related individuals is the ascertainment of a kinship matrix defining the relationship between each pair of individuals in the study.

Traditionally kinship matrices have been estimated based on known pedigree information giving an approximation of the expected level of genetic identity between two individuals based on the number and type of links between them in the pedigree. A simple example of pedigree-based kinship is the case of full siblings with no parental relatedness; At any given point in the genome these siblings have a probability of inheriting the same alleles from both parents of 0.25. The probability of both siblings inheriting only one identical allele is 0.5 with the remaining probability that they each inherit different alleles from both parents, and are therefore genetically unrelated at this position, is 0.25. The pedigree based kinship estimate will therefore be 0.5 in full sibs with no background inbreeding.

Pedigree based methods have two major sources of inaccuracy. The first is the semi random nature of chromosomal recombination means that the theoretical range of average allele sharing for full sibs is between 0 and 1. In practice the observed range is considerably smaller than this but still introduces an element of stochastic

variation that makes pedigree based kinship estimates unrepresentative of the true genetic relationship between samples.

The second source of error in pedigree based kinship estimates is the lack of complete information. Within an artificial animal-breeding framework it may be possible to keep accurate pedigree information over hundred or thousands of generations. In human populations however it may not be possible to trace pedigree relationships more than a few generations due to the long generation interval and a lack of historical records. The founders of populations must therefore be assumed to be unrelated which, particularly in a population isolate, is unlikely to be true. Even where large amounts of pedigree data can be collected the accuracy of the information cannot be guaranteed as it relies on complete reporting of all births and on true paternity always being given which is not always the case.

With the availability of large amounts of genotypic data the relationships between individuals can be estimated more accurately than using pedigree information alone both in the sense of identifying non-paternity through conflicting genotypes and in the sense of gaining a more accurate estimate of the proportion of the genome that a pair of related individuals share.

2.2 MERLIN (multipoint engine for rapid likelihood inference)

The first of the methods used in this project, which utilized pedigree based information, was the calculation of multipoint Identity-By-descent (mIBD) matrices for use in the linkage analysis.

The *MERLIN* program (Abecasis, Cherny et al. 2002) was designed to extend mIBD calculation methods to dense marker sets that could not computationally be handled by previous methods. Calculation of mIBD requires a set of genotypes and pedigree information. Any two individuals can have 0,1 or 2 alleles at a particular locus that appear to be identical by state, IBD estimation attempts to determine the likelihood they are also shared identical by decent through some pedigree pathway that connects both individuals to a recent common ancestor.

For a single marker IBD estimate within a given pedigree the program requires 3 things. Firstly an accurate pedigree going back to a set of unrelated founders. Secondly the genotype data for the individuals whose IBD must be estimated and thirdly an estimate of the founder allele frequencies for that marker.

MERLIN can then calculate all possible sets of founder genotypes and all possible patterns of gene flow through the pedigree that would result in the observed genotypes. In a complex pedigree there will be multiple scenarios through which the observed genotypes could have arisen some of which demonstrate IBD between individuals and some that are IBS. The combined likelihoods of scenarios leading to IBD can then be calculated.

Multipoint methods are used to calculate the likelihood that two individuals within a pedigree are IBD at arbitrarily chosen intervals along the genetic map. This can be

done by extension of the single point IBD calculation to include information from multiple flanking markers and the recombination distances between them.

2.3 SOLAR (Sequential oligogenic linkage analysis routines)

The linkage analysis for this project was carried out using the *SOLAR* package (Almasy and Blangero 1998). The program carries out a variance component analysis to test for evidence that the phenotypic covariance between individuals is caused by their shared genotype at a given genomic location. This analysis was carried out at regular intervals across the genome using the mIBD matrices previously calculated using *MERLIN*.

2.4 Association Analysis

2.4.1 Correcting for population structure

When carrying out a GWAS analysis we expect, and hope for, some deviation of the observed test statistics from the null distribution. The deviation we are hoping to find is a result of some proportion of the SNPs tested being non-randomly associated with the phenotype being tested. There are however several other sources of inflation.

As our samples are taken from isolated populations we expect the greatest source of inflation to arise from increased allele sharing between related individuals and familial covariance in phenotypic values.

Inflation can also results from population stratification whereby two or more genetically divergent populations are treated as a single population. If there are differences in both the allele frequencies and phenotypic distributions of the unobserved sub-populations then false associations will be produced.

A third source of inflation is genotyping error arising from poorly performing genotype assays or errors in the genotype calling algorithm. This can be particularly problematic in case-control study designs when the two sample groups are genotyped separately. In population based studies such as ours the problem is less pronounced as any error should be randomly distributed with respect to the phenotypes being tested.

We can control effectively for the population stratification by conducting individual analysis of each study population we have available and combining the evidence for association using meta-analysis. The issue of family structure is more complex and we have tested a variety of methods to account for this.

2.4.2 Genomic Control

The simplest and fastest method to correct for inflated association statistics in related samples is Genomic Control (Devlin and Roeder 1999). This method is applied after association analysis has been performed and attempts to assess and correct for the overall inflation of test statistics after said inflation has occurred.

Initially the association analysis is performed assuming individuals are unrelated. The association of each SNP with a continuous and normally distributed phenotype is tested by linear regression, along with appropriate covariates and fixed effects.

We then make the assumption that, even for a highly polygenic phenotype, the vast majority of genome-wide markers should not show association and should therefore follow the null distribution. So we take the observed distribution of test statistics, in this case chi-square values with 1 degree of freedom, and find the median value. Subtracting the expected median test statistic gives an estimate of the average level of inflation in our analysis (λ). We can then divide the observed test statistic for all SNPs by λ to obtained the genomically controlled test statistic. The inflation can be visualised with a Quantile-Quantile (QQ) plot in which the observed p-values are plotted against the values expected under the null distribution. Figure 2 shows an example of the uncorrected p-value distribution for height in the ORCADES study, which gives a λ value of 1.44. Figure 3 shows the distribution after correction for this inflation factor. The main body of results now follows the expected null distribution but some degree of inflation is still seen in the tail of the results indicating higher than expected test statistics for a small proportion of SNPs as would be expected if true associated loci are present.

Figure 2. QQ-plot of GWAS results for height in the ORCADES population.

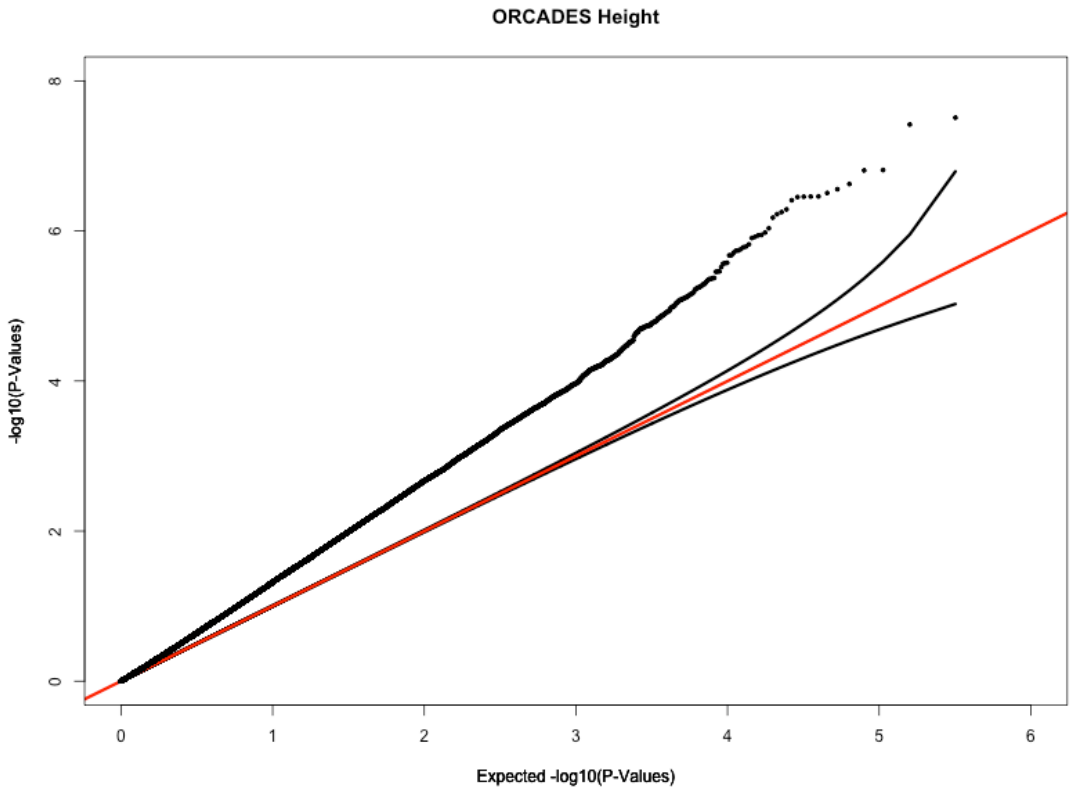
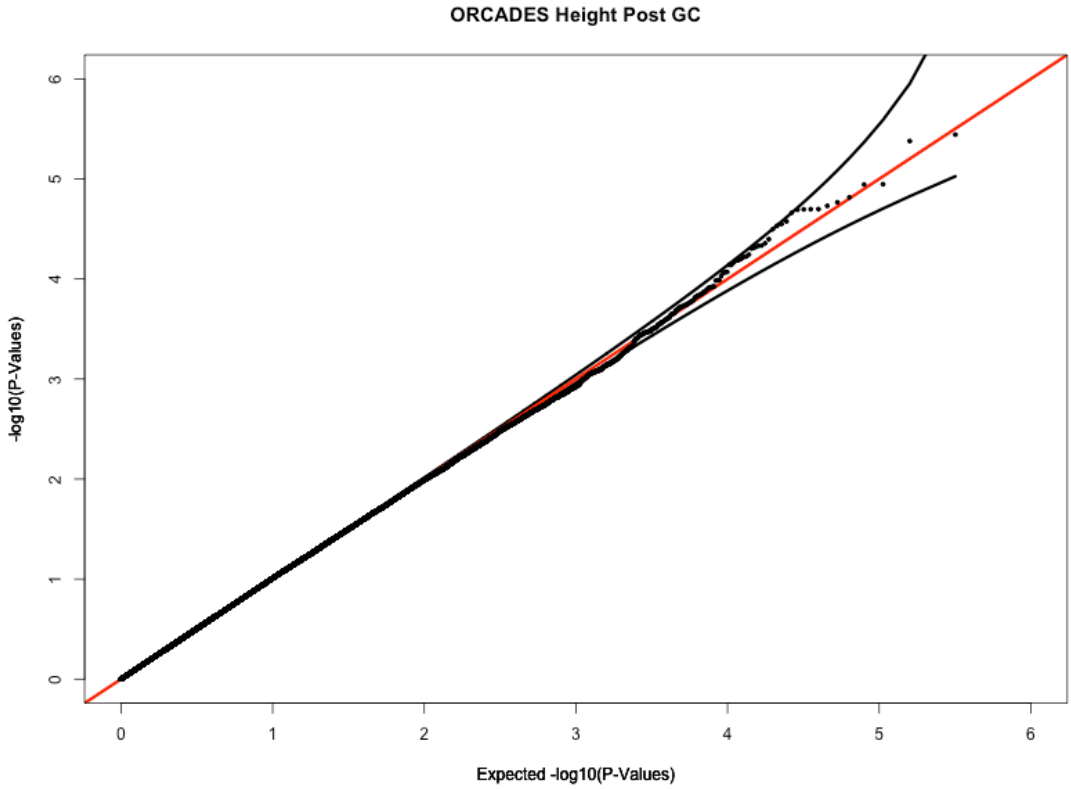


Figure 3. QQ-plot of GWAS results for height in the ORCADES population after correction for inflation by GC (λ =1.44).



The genomic control (GC) method works efficiently when λ is small (Marchini, Cardon et al. 2004), for example when inflation is caused by the background cryptic relatedness that will be present in any population or by some small degree of population admixture. When λ is large, for example due to close family structure within the study, genomic control becomes less anti-conservative and fails to fully correct for inflation (Devlin, Bacanu et al. 2004). Given appropriate pedigree data, alternative methods can be used to more accurately correct for family or population structure prior to association analysis.

2.4.3 GRAMMAR

One of the methods to adjust for relatedness I examined was Genome-wide Rapid Association Using Mixed Model And Regression (GRAMMAR) (Aulchenko, de Koning et al. 2007). GRAMMAR is a multistage method using pedigree information to adjust the phenotype prior to running GWAS.

The first step is to calculate and correct for the proportion of phenotypic covariance that is due to shared kinship, which in this case was done using the ASReml software. The program uses the inverse of the relationship matrix rather than the relationship matrix both because the inverse matrix is sparser and therefore requires less memory to calculate, and because the ultimate aim of the analysis is to remove the effect of the relationship matrix by multiplying out the inverse relationship matrix.

A mixed model can be used to estimate the effects of covariates (i.e. age), fixed effects (i.e. sex) and random effects (i.e. additive polygenic effects and unidentified residual effects). The mixed model takes the form:

$$\gamma_i = \mu + \sum_j \beta_j C_{ji} + G_i + \varpi_i$$

Where γ_i is the phenotype of the *i*th individual, μ is the mean phenotypic value, β_j is the estimated effect of the *j*th covariate or fixed effect, C_{ji} is the value of the *j*th

covariate or fixed effect for the *i*th individual. G_i is the additive polygenic effect and ϖ_i is the residual random effect that we wish to calculate.

By rearrangement the residual phenotypic value for the *i*th individual can be given by:

$$\varpi_i = \gamma_i - (\mu + \sum_j \beta_j C_{ji} + G_i)$$

G_i is defined as the kinship, or relationship, matrix multiplied by the additive polygenic variance or estimated narrow-sense heritability of the phenotype.

These residual random effect values can then be used in place of the original phenotypic value in a linear regression against SNP genotype and other appropriate covariates and fixed effects:

$$\varpi_i = \mu + kg_i + e_i$$

Where g_i is a vector of genotypes at a given SNP k is the effect of that SNP and e_i is the remaining residual error.

The GRAMMAR method should perform better than GC when heritability estimates are substantial and a large degree of inflation is expected. It does however rely on the accuracy and completeness of pedigree information.

2.4.4 FASTA

The third and final analysis method used to adjust for relatedness within the sample was based on the Family Based Score test for Association (FASTA) (Chen and Abecasis 2007). The method is similar to GRAMMAR in the sense that both methods use an initial step to estimate the polygenic covariance and a second step to test for association. FASTA differs from GRAMMAR in that having estimated the polygenic and residual variance both are used, along with the phenotype in calculation of the association test statistic.

Rather than using a pedigree based kinship matrix, genomic kinship matrices were calculated for participants within each study using identity-by-state sharing, weighted by allele frequency. The kinship matrix was then used in a linear polygenic mixed model to account to estimate the polygenic and residual environmental components.

The association test statistic is then calculated according to the formula:

$$T_F^2 = \frac{((g - E[g])T \cdot (\Phi \cdot \hat{\sigma}_g^2 + I \cdot \hat{\sigma}_e^2)^{-1} \cdot (Y - \hat{\mu}))^2}{(g - E[g])T \cdot (\Phi \cdot \hat{\sigma}_g^2 + I \cdot \hat{\sigma}_e^2)^{-1} \cdot (g - E[g])}$$

Where *(g - E[g])* is the additively coded observed genotype minus expected average genotype based on allele frequency. $(\Phi.\sigma_g^2 + I.\ \sigma_e^2)^{-1}$ represents the inverse of the variance covariance matrix calculated by multiplication of the kinship matrix, Φ, by the estimated polygenic variance, σ_g^2, and adding the residual environmental

variance, σ_e^2, multiplied by the identity matrix, *I*. *Y* is the phenotype value and μ is the population mean.

The FASTA method should provide suitable correction across a wide range of λ values and, unlike GRAMMAR; it does not rely on pedigree information. This makes it more versatile in that it can be applied to populations with unknown pedigree information. It should also theoretically give more accurate correction in that human pedigrees can rarely be traced beyond a handful of generations with any degree of accuracy, whereas the genomic kinship should always give an accurate reflection of the genetic relationship between two individuals.

2.5 Meta-Analysis

Meta-analysis of the association statistics was performed using an inverse variance method implemented in the MetABEL package (Aulchenko, Ripke et al. 2007).

$$\omega_{ij} = \frac{1}{S_{\cdot ij}^{\ 2}}$$

$$\hat{\beta}_i = \frac{\sum_{1}^{j} \beta_i \cdot \omega_i}{\sum_{1}^{j} \omega_i}$$

$$\hat{S}_i = \frac{1}{\sum_{1}^{j} \omega_i}$$

Where β_{ij} is the effect estimate for SNP i in study j, S_{ij}^2 is the standard error squared and ω_{ij} is the study weighting. The meta-analysis effect estimate for each SNP is then calculated by multiplying each study estimate by its weighting, summing across all studies and dividing by the sum of all study weights. The meta-analysis standard error ($\hat{S}_i$) is given by 1 over the sum of all study weights.

2.6 Accounting for medication effects

When dealing with quantitative phenotypes certain medications can alter the phenotype in question. For example individuals with T2D who are unable to maintain dietary control of their blood glucose levels may be prescribed glucose lowering medications(Consoli, Gomis et al. 2004) or, in more extreme cases, insulin injections(Mayfield and White 2004). These medications will have a substantial impact on the measured phenotype and would therefore interfere with the accuracy of our estimates of other sources of variation such as genetic loci.

A variety of strategies are available to deal with the effects of medication roughly divided into three categories; i) ignoring the medication, ii) attempting some form of correction to restore the phenotypic value to the untreated level, iii) excluding medicated individuals prior to analyses.

The category of correcting for medication effects includes strategies in which the phenotype value of all medicated individuals is changed by some constant value or alternatively a statistical correction which allows for a varying level of treatment effect across in different individuals.

A comprehensive review of strategies to deal with medication has been presented in the work of Tobin et al (Tobin, Sheehan et al. 2005). The recommendation of this work favour correction of phenotypic value either through addition of a constant value or using a censored normal regression model to estimate a different correction value for each medicated individual depending on their position in the distribution.

The analysis conducted on glycaemic phenotypes in this thesis excludes all medicated and hyperglycaemic individuals. The method of adding a constant value to medicated individuals was considered, however treatment for type 2 diabetes can vary from diet and exercise in well controlled cases (Boule, Haddad et al. 2001), oral glucose lowering drugs in more severe cases(Consoli, Gomis et al. 2004; Krentz and Bailey 2005), or even insulin treatment in poorly controlled cases(Mayfield and White 2004). Highly detailed information of this type was not readily available from all of the studies used in this project and so this method of correction was not feasible.

The strategy of estimating and correcting for medication effects statistically was also considered. There are several important considerations when using this approach

including the size of the sample available, the proportion of medicated individuals and the distribution of the phenotype. The censored normal regression method assumes that

> *"the distribution of underlying BP above any specified value is the same in treated and untreated individuals"*

While this may be a valid assumption in relation to blood pressure the distribution of diabetes related phenotypes in our study populations was highly skewed with diabetic patient values being considerably higher than the normal range. The overall sample sizes in our studies were modest and the prevalence of diabetes in our populations, between 2.1 and 9.2%,was also relatively small compared with more widely treated phenotypes such as blood pressure. All these factors result in a very small amount of informative data for the censored normal regression correction method.

In addition to data driven arguments the decision to exclude medicated and hyperglycaemic individuals from our analysis was also guided by the strategy of previous studies. The exclusion of known diabetics in studies of related quantitative phenotypes has become common practice(Meigs, Manning et al. 2007; Chen, Erdos et al. 2008; Pare, Chasman et al. 2008; Bouatia-Naji, Bonnefond et al. 2009; Prokopenko, Langenberg et al. 2009; Dupuis, Langenberg et al. 2010).

3 Chapter 3: Linkage meta-analysis of fasting glucose in four isolated populations

The first analysis attempted in this project was a genome-wide linkage analysis of Fasting Glucose (FG) levels in the four EUROSPAN populations described below with appropriate marker and phenotype data available. Blood glucose levels vary greatly in response to the intake of food and some beverages so FG measures taken after an overnight fast are a more stable measure of baseline blood glucose.

Generally speaking, highly polygenic quantitative phenotypes may not be expected to yield strong results using the linkage analysis method since, as previously noted, the method has weak power to detect variants with small individual effect sizes. However, the existence of at least 10 forms of monogenic diabetes (Craig, Hattersley et al. 2006) may provide more reason for hope. While the mutations responsible for monogenic forms result in early onset of diabetes and the related breakdown of glycaemic control, alternate polymorphisms at the same loci may produce less severe functional consequences, which on their own are not fully penetrant but may contribute to the development of T2D in concert with other genetic or environmental factors. A second, practical reason for conducting linkage analysis is that it makes use of the large amounts of family structure inherent in any isolated population study.

The four populations were tested for linkage separately and the test statistics were combined using meta-analysis to give an overall measure of evidence for linkage. Regions that showed suggestive evidence of linkage for FG, either in a single population or in the meta-analysis, were subsequently tested in a candidate region association analysis using high density SNP data.

The theoretical rationale for this candidate region analysis was that; while linkage analysis is most powerful in the detection of rare, strong variants and association analysis is most powerful for detecting common variants, a true FG locus may show marginal evidence in both analyses. On the other hand a false positive finding arising from some problem with the linkage markers would be unlikely to show independent association evidence in the SNP data.

3.1 Description of populations/markers for ERF, MICROS, NSPHS and VIS Populations

Four of the EUROSPAN populations originating from Sweden, Italy, Croatia and the Netherlands had genotype data designed for the purposes of linkage analysis.

The Swedish samples are part of the Northern Swedish Population Health Study (NSPHS) representing a family-based population study including a comprehensive health investigation and collection of data on family structure, lifestyle, diet, medical

history and samples for laboratory analyses (Johansson, Vavruch-Nilsson et al. 2005). The samples used for linkage analysis were collected from the southern part of the Swedish mountain region in the county of Västerbotten. Historic population accounts shows that there has been little immigration or other dramatic population changes in this area during the last 200 years. In total 436 individuals were genotyped for 390 microsatellite markers.

The Italian population subjects were sampled from a study of microisolates in south Tyrol (MICROS). Study participants were volunteers from three isolated, German-speaking villages, located in a region bordering with Austria and Switzerland (Pattaro, Marroni et al. 2007). Due to geographical, historical and political reasons, the entire region experienced prolonged isolation from surrounding populations. The investigated population is characterized by an old settlement, a small number of founders, high endogamy rates, slow or null population expansion and negligible immigration. An extensive genealogy spanning 12 generations was available for this study. A total of 926 individuals from this population were genotyped at 1,113 microsatellite markers.

The Croatia population (VIS) consists of unselected Croatians aged 18–93 years recruited into the study from the villages of Vis and Komiza on the Dalmatian island of Vis who were phenotyped for >50 disease-related quantitative traits. The genetic isolation of the villages from the Croatian mainland and from surrounding islands has been confirmed (Vitart, Biloglav et al. 2006) and this population has already

yielded novel and replicated quantitative trait loci(Vitart, Rudan et al. 2008). 747 microsatellites were genotyped in 591 members of the VIS study.

The Dutch samples are from the Erasmus Rucphen Family (ERF) study, which was carried out on a Dutch isolated population located in the Southwest of the Netherlands. This community was founded in the middle of the 18th century by approximately 150 individuals and was isolated until the last decades. It is characterized by rapid growth and minimal inward migration and has now expanded up to 20,000 inhabitants (Aulchenko, Heutink et al. 2004). Within this population, a specific subpopulation based on 20 couples (selected on the basis that they had at least 6 children baptized in the community church between 1880 and 1900) has been defined. All living descendants of the selected couples and their spouses (n≈3,000) have been recruited, basically forming one single extended pedigree. The ERF study samples were typed at a set of 6,008 SNPs designed for genome wide linkage analysis. A total of 1,455 ERF samples were genotyped making this by far the largest population.

The use of different genotype panels in these studies raises several issues. The considerably large number of markers used in the ERF study is indicative of the reduced information content of SNP markers when compared to microsatellites. SNPs are predominantly bi-allelic and, even in cases when 3 or 4 alleles exist; the duel fluorescence microarray assay used to type them assumes this bi-allelic state.

By contrast the microsatellite markers used in the other studies have up to 14 distinct alleles making the differentiation of IBS and IBD considerably easier.

A second issue that arose within the microsatellite-based analyses was the genetic map position of markers. The linkage analysis requires marker positions to be assigned according to recombination distances between markers rather than physical base-pair position. These recombination distances are calculated within a population by observation of crossover events within pedigrees and; since larger studies will give more accurate estimates of recombination distance, a number of standardized recombination maps have been published(Murray, Buetow et al. 1994; Dib, Faure et al. 1996; Broman, Murray et al. 1998; Kong, Gudbjartsson et al. 2002).

The VIS and MICROS marker positions were based on the Decode recombination map(Kong, Gudbjartsson et al. 2002) while the NSPHS markers positions were assigned according to the older Marshfield map (Broman, Murray et al. 1998). To converge all marker data to the same genetic map we replaced the Marshfield positions with Decode positions in the NSHPS data wherever a marker was reported on both maps. In cases where a marker was not present on the Decode map we used the two nearest flanking markers for which Decode and Marshfield positions were available to interpolate an approximate Decode position.

3.2 Pedigree Splitting

Large pedigrees are desirable when conducting a linkage analysis, however the computational requirements for calculating multipoint IBD (mIBD) estimates increases exponentially with pedigree size. The size and complexity of the pedigrees obtained for the EUROSPAN populations was such that mIBD estimates could not be obtained with the resources available at the time. To overcome this problem the pedigrees were split into multiple sub-pedigrees using the *PedStr* software (Kirichenko, Belonogova et al. 2009) with a maximum bit size of 18 individuals, where bit size is twice the number of non-founders minus the number of founders within the sub-pedigree.

3.3 IBD calculations using multipoint engine for rapid likelihood inference (MERLIN)

Having obtained sub-pedigrees of an appropriate size the mIBD matrices were calculated for each population using the *MERLIN* software (Abecasis, Cherny et al. 2002). Splitting the pedigrees gives a slightly reduced estimate of pedigree-based kinship between individuals, which will in turn cause an underestimation of mIBD. To correct for this we calculated single marker IBD estimates, took a genome-wide average and used this as an inflation factor for the mIBD estimates.

3.4 Phenotype

FG levels were available for the majority of samples within each of the 4 studies described. Individuals with known diabetes diagnoses or who had FG values above 7mmol/l were excluded from analysis so that FG variation in the healthy population could be tested.

The proportion of samples excluded for diagnosed or suspected diabetes varied considerably between the four studies with a maximum of 7.6 % from the ERF study and minimum of 2% from the NSPHS study. Within the non-diabetic populations there was still a great deal of variation in the distributions of relevant measures. Table 1 shows the mean and standard deviations for age, BMI and FG levels in the four populations. This phenotypic heterogeneity coupled with the genetic heterogeneity between separate populations is the major rationale for using a meta-analysis rather than pooled analysis study design.

Table 1. Descriptive statistics for populations used in linkage analysis, mean (standard deviation)

Population	N	Age (yrs)	BMI (kg/m^2)	FG (mmol/l)
ERF	1455	52 (16.0)	26.8 (4.6)	4.6 (1.02)
MICROS	926	46 (16.6)	25.6 (4.5)	4.7 (0.92)
VIS	591	56 (16.3)	27.1 (4.1)	5.7 (1.48)
NSPHS	436	45 (11.6)	25.8 (3.8)	5.1 (1.17)

3.5 Linkage analysis using Sequential Oligogenic Linkage Analysis Routines (SOLAR)

The estimation of trait heritability and linkage analysis was performed using the *SOLAR* software package (Almasy and Blangero 1998). This package has its own methods for calculating mIBD matrices but, while computationally efficient for large pedigrees, the mIBD values obtained are approximate. The alternative *MERLIN* algorithm we chose to use gives exact mIBD values. Although using split pedigrees is itself an approximation to the true underlying IBD, the MERLIN method applied to split pedigrees captures all close familial relationships while running considerably faster than SOLAR.

The *SOLAR* program outputs an estimated heritability for the trait being analyzed and a list of LOD-score test statistics denoting evidence for linkage at the specified map positions. The meta-analysis test statistic was simply the sum of the individual study statistics at each position. For the purposes of assessing significance in the meta-analysis it is important to note that the *SOLAR* linkage analysis constrains the output LOD-score to a minimum of zero, i.e. no evidence against linkage is given. This results in a general inflation of the meta-analysis results and so requires a higher significance threshold.

To define our significance thresholds we used the relationship that a LOD score multiplied by twice the natural log of 10 ($2x\log_e(10)$) is approximately equivalent to a χ^2 value (Lander and Kruglyak 1995). Using a p-value of $5x10^{-4}$ and 1 degree of

freedom gives a LOD threshold for a single study of 2.63. For the meta-analysis the same p-value with 4 degrees of freedom gives a threshold of 4.34.

3.6 Results

The narrow sense heritability estimate for each study calculated using *SOLAR* is given in Table 2. In terms of accuracy the larger studies would be expected to give estimates closer to the true genetic heritability. Within our study populations the two largest studies, ERF (N=1455) and MICROS (N=926), give the smallest (0.3) and largest (0.5) estimates respectively. A previous extremely large study of a Sardinian population isolate gave an estimated serum glucose heritability of 0.36 (Pilia, Chen et al. 2006) consistent with the ranges we see here. As previously mentioned the heritability estimates are population specific and there are several possible explanations for the wide range of estimates obtained for these four studies.

Firstly environmental variance will be affected by the variability, or lack thereof, of diet and physical activity within each study population. Secondly the accuracy of the measurement method used in each study site and the consistency of the measurement protocol will result in variable levels of error variance for each study. Thirdly there may theoretically be polymorphic loci that influence FG levels segregating in one or more of the study populations that are fixed in the other populations. It is most likely that a combination of these factors is responsible for the variable estimates but the most important fact is that all studies show a substantial genetic component for FG levels indicating that there are indeed loci to be identified.

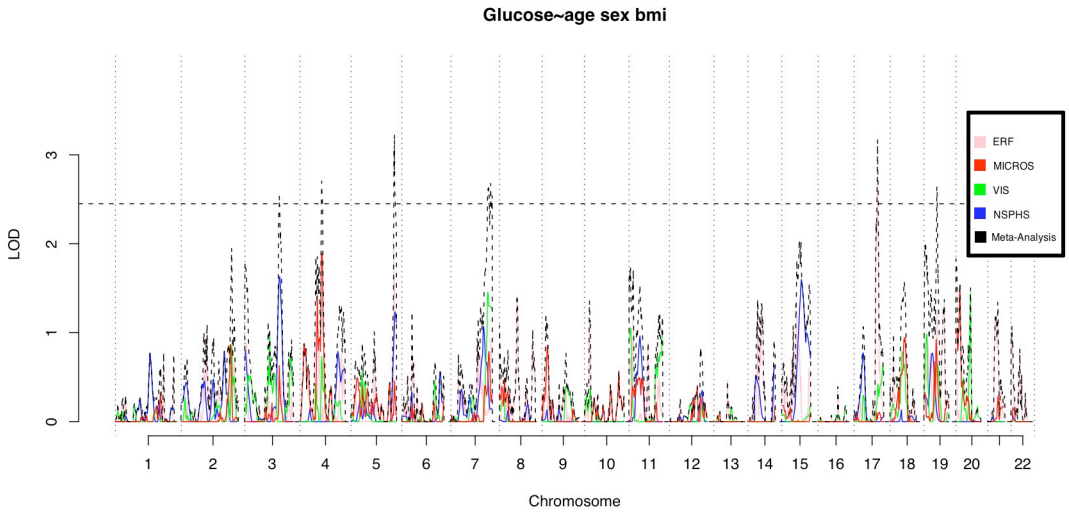
Table 2. Narrow sense heritability estimates for FG levels using SOLAR

Population	N	FG mmol/l mean (SD)	Heritability (h^2)
ERF	1455	4.6 (1.02)	0.30
MICROS	926	4.7 (0.92)	0.50
VIS	591	5.7 (1.48)	0.41
NSPHS	436	5.1 (1.17)	0.43

The genome wide multipoint linkage analysis results for each of the four study populations individually and the combined meta-analysis result are shown in figure 4. The results from single population analysis show a single peak that reaches the threshold for significance. The peak, on the q arm of chromosome 17, is significant in the ERF population while being almost completely absent in the other three study populations.

Examination of the meta-analysis results shows no statistically significant results but several other peaks on chromosomes 3,4,5,7 and 19 were included in subsequent analysis because they contained modest but consistent linkage signals from multiple populations.

Figure 4. Genome-wide linkage meta-analysis for FG in 4 populations



For the purposes of candidate region association analysis we defined linkage peaks by subtracting 2 LOD points from the highest value of a peak and including the flanking regions with scores above this value. The size of each of the selected regions and the number of genotyped SNPs are shown in table 3.

Table 3. Linkage regions for FG and the number of SNP markers available within each peak for candidate region association.

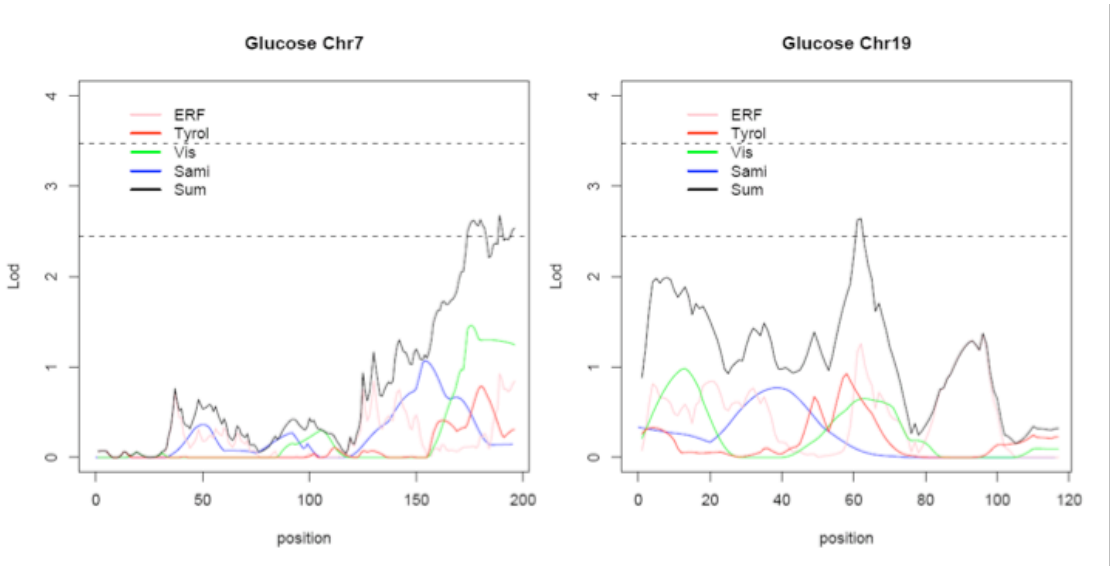
Chromosome	Peak Position (cM)	Width (cM)	Number of SNPs
3	162	48	5287
4	102	20	2506
5	205	16	752
7	188	62	3178
17	111	18	1287
19	61	21	1448

Collaborators from each study group conducted genotypic association tests for the SNPs identified in each candidate region. The total number of SNPs tested in this candidate region analysis was 14,458 giving a significance threshold of $3.5x10^{-6}$. The majority of regions tested showed no evidence for association at or close to this significance threshold in any of the individual studies. The peak on chromosome 19 showed strong association with marker rs11668501 in the ERF study along with weaker association with several flanking markers as shown in table 4.

Table 4. Chromosome 19 linkage peak candidate region association results in ERF study.

SNP	Chr	Position (mb)	Effect AB	Effect BB	P-value	MAF
rs3865452	19	45.90	-0.04	0.13	0.01	0.45
rs2604913	19	45.91	-0.05	0.15	0.0022	0.42
rs890934	19	45.92	-0.05	-0.04	0.01	0.41
rs11668501	19	45.92	-0.02	1.06	$1.2x10^{-8}$	0.12
rs3745216	19	45.94	-0.03	0.13	0.01	0.44

Figure 5. Linkage meta-analysis results for FG in 4 populations showing previously reporting linkage peaks.



3.7 Discussion

Two of the regions identified in this study have been previously identified in linkage studies of glucose levels or T2D status (Figure 5). A comparable meta-analysis found significant linkage for fasting glucose on chromosomes 7q36 and 19q13 (An, Freedman et al.), which overlap with the peaks found in our study. The q-arm of chromosome 7 has also shown suggestive evidence of linkage for T2D in Pima Indians (Imperatore, Hanson et al. 1998).

No significant linkage evidence was found on chromosome 10q in the region of the TCF7L2 gene, which has been identified in many independent populations, both by linkage and association analysis, as a significant risk locus for T2D.

The most significant linkage result, on chromosome 17q in the ERF study, has not shown consistent evidence in previous studies of T2D of fasting glucose levels however it has been implicated in the rarer monogenic form of diabetes Maturity Onset Diabetes of the Young (MODY) and designated MODY5. There are a number of candidate genes in this region that could affect glucose homeostasis including *TCF2*.

An alternative approach not taken forward in this project is the analysis of phenotypic extremes as a binary phenotype. Our chosen strategy involved the removal of medicated diabetics and observed hyperglycaemic individuals. In an extreme analysis strategy these individuals would be grouped together as a set of cases. A second subset of the population can then be taken from the opposite end of the phenotypic distribution to act as controls.

Selecting phenotypic extremes has been proposed as should provide enrichment for alleles with large effects (Carey and Williamson 1991). This is particularly useful when applied prior to genotyping as the phenotypically selected sample can provide greater power than an unselected sample of equivalent size(Allison 1997). In the case of our studies pre-selection of samples for genotyping based on phenotype was not

practical as the studies were designed to investigate a wide range of traits. Given this fact an analysis of phenotypic extremes would allow us to retain the diabetic and hypoglycaemic samples but would simultaneously remove the information provided by genotyped samples in the middle of the phenotypic distribution.

4 Chapter 4: Genome-wide association of glycaemic traits in three to five isolated populations

4.1 Background on hyperglycaemia/Type II diabetes genetics

Diabetes is characterized by one or more of impaired fasting glucose, impaired glucose tolerance, defective insulin secretion or defective insulin action. T2D is a common complex disease characterized by impaired fasting glucose (IFG) and/or impaired glucose tolerance (IGT) as defined using criteria endorsed by the world health organization (WHO 2006). A fasting plasma glucose level above 7mmol/l on two separate occasions is considered diagnostic for T2D. Chronic hyperglycaemia can lead to serious complications including nephropathy, neuropathy, retinopathy and increased risk of cardiovascular disease (Grant, Thorleifsson et al. 2006). The total number of diabetes sufferers worldwide was estimated at over 171 million for the year 2000 and is predicted to rise to 366 million by the year 2030(Wild, Roglic et al. 2004).

Environmental factors known to influence T2D risk include central obesity and a lack of physical exercise. Recent increases in the prevalence of T2D are likely to be due to the increased prevalence of adverse lifestyle factors, ageing of populations and improved survival, however T2D also shows a significant genetic component with a sibling risk ratio of 3.5 (Weijnen, Rich et al. 2002). It has been suggested that

genetic mapping of a disease as complex and heterogeneous as T2D is inappropriate and that study of the quantitative traits underlying the disease may be more productive(Bougneres 2003). However in recent years studies of T2D using case control designs have successfully identified and replicated 37 T2D susceptibility loci demonstrating that, with sufficiently powered studies, the genetics of heterogeneous diseases and phenotypes can be studied in this way (Sladek, Rocheleau et al. 2007; Ng, Park et al. 2008; Zeggini, Scott et al. 2008; Bouatia-Naji, Bonnefond et al. 2009; Voight, Scott et al. 2010).

Several biochemical traits are relevant to T2D. FG levels are commonly used as a diagnostic tool in suspected T2D cases with levels above 7 mmol/L being suggestive of T2D. Glycosylated haemoglobin (HbA_{1C}) is used to monitor blood glucose levels over the complete life cycle of the red-blood cells, approximately 8-12 weeks. This makes the trait ideal for monitoring the health of T2D patients and it is also increasingly used as a diagnostic test for T2D to obviate the need to collect a fasting sample or administer a 2-hour glucose challenge test. Insulin levels are also a key component of glucose homeostasis and can be used along with FG levels to compute the Homeostasis Model Assessment phenotypes HOMA-IR (Insulin Resistance) and HOMA-B (Beta-cell function) (Matthews, Hosker et al. 1985).

One of the first major genetic risk factors to be discovered for T2D is the *TCF7L2* locus that was identified through association mapping within a region previously identified through linkage on chromosome 10q1. This association has been

consistently replicated across a diverse range of populations (Kimber, Doney et al. 2007; Sladek, Rocheleau et al. 2007; Herder, Rathmann et al. 2008; Ng, Park et al. 2008; Roth, Hinney et al. 2008; Sanghera, Ortega et al. 2008).

More recently high density Single Nucleotide Polymorphism (SNP) genotyping has allowed the development of genome-wide association (GWA) strategies, which can identify QTLs in linkage disequilibrium (LD) with a SNP marker within a population rather than relying on segregation of markers and QTLs within a family. Whilst linkage analysis is useful for detecting rare variants with large effects the GWA approach can detect more common variants with smaller individual effect sizes.

Many GWA scans have been carried out for T2D with considerable success, and at least 38 reproducible associations with small individual effect sizes have been identified(Lango, Palmer et al. 2008; Voight, Scott et al. 2010). Additionally as people recognise the benefits of studying quantitative traits a number of GWA scans have begun to study glucose, insulin and HbA_{1C} as quantitative traits.

The Framingham heart study published one of the first GWA studies on diabetes related quantitative traits and were unable to identify any associations which reached a genome-wide significance level after correcting for the number of tests performed(Meigs, Manning et al. 2007). However the study had a limited sample size of just over 1000 individuals and a relatively sparse 100k SNP marker panel

which, in light of the effect size of variants now known to be involved in these phenotypes, was greatly underpowered.

More recently collaborative efforts between many smaller studies have produced convincing, replicated associations. Chen et al (2008) used an initial sample of 5k non-diabetic Finnish and Sardinian individuals and a replication sample of over 18k European individuals to study FG levels (Chen, Erdos et al. 2008). They identified a region between the *G6PC2* and *ABCB11* genes that was strongly associated ($p=6.4 \times 10^{-33}$) with FG.

A later study using French non-diabetic individuals identified the *MTNR1B* melatonin receptor gene as having a role in FG levels (Bouatia-Naji, Bonnefond et al. 2009). This study had a relatively small discovery cohort of 2,151 samples and produced two significant associations after correction for multiple testing. Only one of the markers replicated consistently across four replication cohorts totaling over 16 thousand samples suggesting that the other result may have been a false positive despite having a considerably more significant p-value (8×10^{-9} and 1.3×10^{-7} respectively).

The most recently, and to date the largest, published GWA meta-analysis study conducted on FG was by the Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC) (Prokopenko, Langenberg et al. 2009). The first stage of this collaboration included a total of 33,321 healthy individuals of European descent and

2,688 diabetics. The study provided extremely strong replication of the previously identified effects of *MTNR1B*, *G6PC2* and the gluco-kinase receptor (*GCK*) on FG levels. The study also looked at the effect of identified FG risk alleles on T2D risk but found no significant effect. A second phase of the MAGIC collaboration was published during the writing of this thesis and includes analysis of data for over 109 thousand individuals (Dupuis, Langenberg et al. 2010).

Studies with such large sample sizes should be capable of detecting the common variants with small effect sizes that are predicted to make up the majority of genetic variance in complex disorders. It is however still possible that a rare variant which may be at a low frequency, or non-polymorphic, in some populations could be identified in an isolated population where it may have drifted to a higher frequency. A number of the studies involved in the collaborative meta-analyses were designed for just that purpose. The problem in these cases is that replication becomes more difficult and if an association fails to replicate it is not usually pursued any further.

The isolates used in this study have had small stable population sizes for long periods resulting in reduced allelic heterogeneity and stronger linkage disequilibrium than is found in cosmopolitan populations (Wright, Carothers et al. 1999). The reduced genetic and environmental variance in populations of this type should improve identification of variants with small effect sizes.

4.2 Description of populations/SNP markers for ERF, MICROS, Vis, KORČULA, ORCADES

The data sets used for association analysis partially overlapped with the data sets previously used in the linkage analysis with the MICROS, ERF and VIS samples used for both linkage and association analysis. One additional EUROSPAN project, the Orkney Complex Disease Study (ORCADES), was added to the analysis at the GWAS stage and a further population isolate from the Dalmatian island of Korčula (KORČULA) was available for association analysis.

The ORCADES project is an ongoing family-based cross-sectional study in the isolated Scottish archipelago of Orkney. Genetic diversity in this population is decreased compared to Mainland Scotland, consistent with the high levels of endogamy historically. Data for participants aged 18-100 years with ancestry from a subgroup of ten islands, were used for this analysis. Fasting blood samples were collected and over 200 health-related phenotypes and environmental exposures were measured in each individual. All participants gave informed consent and the study was approved by Research Ethics Committees in Orkney and Aberdeen. A total of 719 ORCADES samples passed quality control and were used in the association analysis.

The VIS study as previously described consists of samples from two villages on the island of Vis with known reduced genetic diversity. A larger number of samples

were genotyped for association analysis and, after quality control, a maximum of 795 samples were used.

The ERF samples typed with genome wide SNP coverage represent a subset of those used for linkage analysis. After quality control a maximum of 918 samples were available for association analysis.

The MICROS samples used in the association analysis are also from the same study population used in the linkage analysis. A total of 1097 samples had genome-wide SNP data.

The KORČULA project followed on from the VIS project using a similar study design to collect samples from a second Dalmatian island, Korčula. A vast range of anthropometric, biochemical and psychometric measurement were collected as well as information on medical history and various lifestyle factors with relevance to one of more of the measured phenotypes. Quality controlled genotype data was available for a total of 888 KORČULA samples.

4.3 Quality Control of Genome-wide SNP data

Proper Quality Control (QC) of large-scale SNP genotype data is important for avoiding large numbers of false-positive and false negative association results that may arise due to errors in genotyping or the presence of population

stratification(Cardon and Palmer 2003; Anderson, Pettersson et al. 2010; Weale 2010). Accuracy of genotype data is also of relevance to down-stream data applications such as imputation wherein a badly called SNP may be used to infer information about hundreds of additional untyped variants, although imputation procedures have been shown to be reasonably robust with respect to standard QC protocols(Southam, Panoutsopoulou et al. 2011).

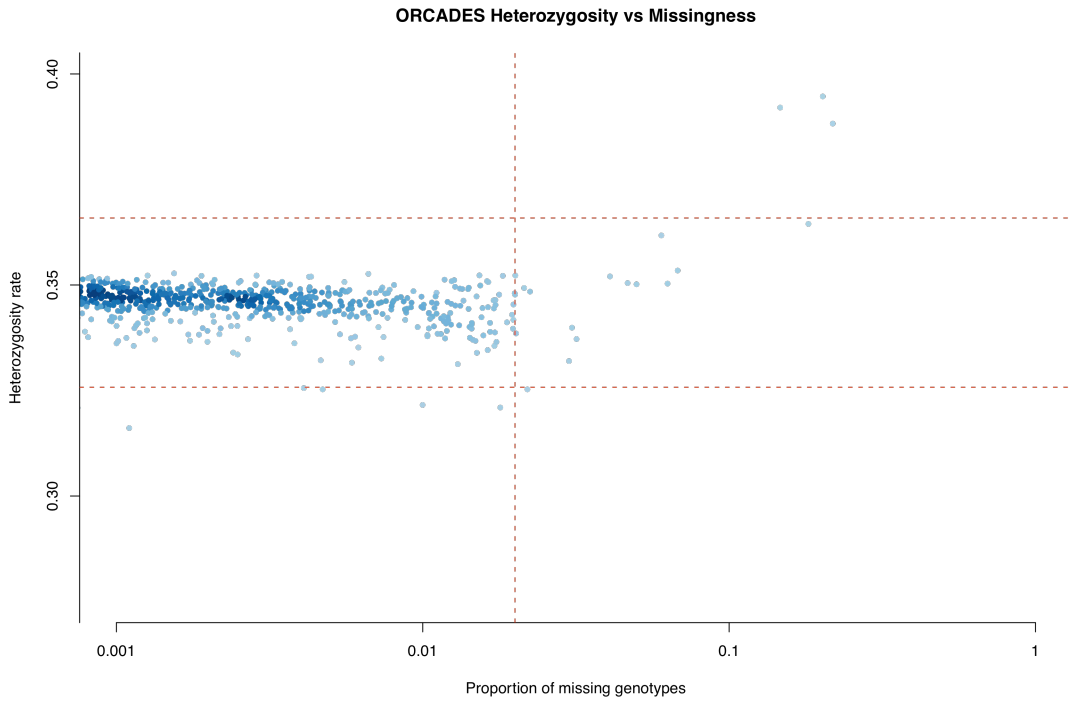
The initial wave of data for the ORCADES project provided genotype data for a total of 761 samples. The first check performed on this data was to assess the genetic sex of each individual's X chromosome SNP data to calculate an inbreeding coefficient F. In females the X chromosome F should be close to the coefficient for autosomal data, though slightly higher due to the reduced effective population size for the X chromosome. As males only carry a single X chromosome all X-linked SNPs should appear homozygous in males resulting in an inbreeding coefficient of 1(Weale 2010). Having obtained a predicted genetic sex we then cross-reference this with the reported sex for that sample. While there are circumstances, such as X chromosome aneuploidy or gender reassignment, that could result in the reported and genetic genders conflicting it is more likely in these cases that mislabeling of a DNA sample or some other identifier error has occurred and the genotype sample and phenotypic data cannot be confidently linked to one another. So in these gender mismatch scenarios the sample must be excluded from further analysis. This sex-check excluded 9 samples from the ORCADES data set.

Individuals with poor quality genotype data can be identified by their genotype call rate and/or their average heterozygosity as shown in figure 6. An individual with low average call rate may represent a low concentration or degraded DNA sample and the genotypes that are called in these individuals may therefore be unreliable.

The average genome-wide heterozygosity of a sample can identify unusual samples in a number of ways. Individuals with unusually high heterozygosity may result from population admixture and so would be unrepresentative of the population being studied. Alternatively extremely high heterozygosity may result from contamination of the DNA sample, which clearly means the genotypes are not accurate(Weale 2010). In the opposite instance of unusually low heterozygosity the explanation is likely to be either increased parental relatedness of the individual resulting from past inbreeding events and a proportionally inflated level of IBD across the genome or poor quality DNA.

To define and exclude samples that may adversely affect analysis we used a genotype call rate of 98% as shown by the vertical line in figure 6, 27 individuals fell below this threshold. Within these 27 samples, 2 also had heterozygosity more than 3s.d above the population mean and were excluded. Individuals with good call rate but unusually low heterozygosity were not excluded because the use of an isolated population increases the levels of background inbreeding and so we expect these low heterozygosity values to be the result of increased parental relatedness rather than poor quality DNA samples.

Figure 6. Individual call rate and heterozygosity values for ORCADES individuals retained in association analyses. Vertical line represents call rate of 0.98. Horizontal lines are +/- 3s.d from population mean heterozygosity.



Association analysis is also sensitive to the relationships between samples and in particular twin or duplicated samples can introduce false positive results. Identification of genetically identical samples was performed by calculating the identity by state (IBS) levels of a random subset of 2000 autosomal SNP markers and identifying pairs of samples that had excessively high allele sharing at these markers. The mean IBS in the ORCADES samples was 0.72 (s.e. 0.009) and a threshold of 0.95 was used to identify a total of 3 pairs of duplicate samples. In each of these cases one of the pair of individuals had already been marked for exclusion based on call rate.

Lastly we identified samples that were genetically too distant from the majority of the population. This was initially carried out by multidimensional scaling (MDS) of IBS distances. The majority of samples within a population should form a single cluster while samples with different ethnic backgrounds will have consistently lower IBS and will stand out from the rest of the population. The MDS analysis identified a further 6 individuals for exclusion.

4.3.1 Principle Components Analysis (PCA) of study populations

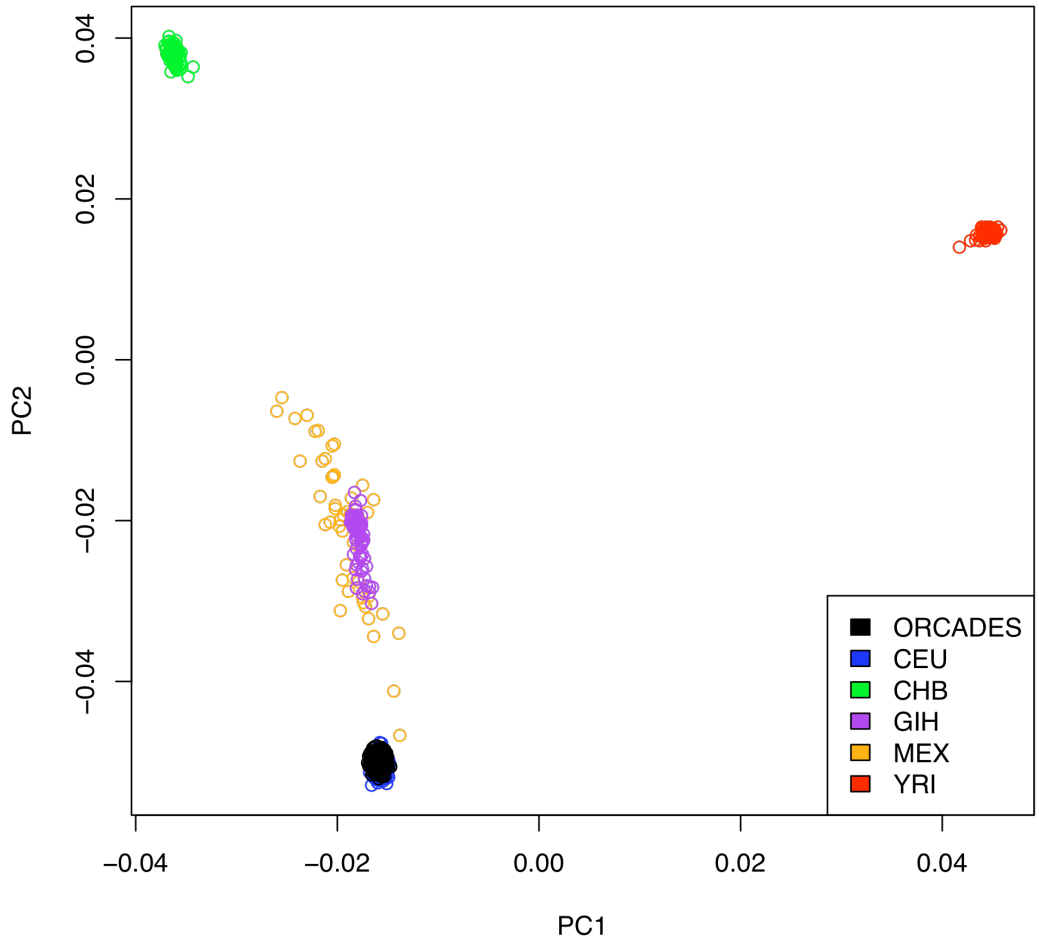
To confirm the quality control results obtained with simple MDS we subsequently analyzed the clean data set using more versatile PCA methods. This was done firstly to confirm the European ancestry of our ORCADES samples and subsequently, when moving into meta-analysis, to explore the finer scale ethnic differences between the ORCADES population and the other study populations from the EUROSPAN consortium and the KORČULA project

For the first broad scale ethnic check we used the phase 3 samples from the HapMap project (2003) as an ethnically diverse training data set for the PCA analysis. This means that the weightings for the SNPs that make up the major principle components (PCs) are chosen based on differences between the ethnically diverse HapMap populations rather than our own populations. The construction of the PCs is sensitive to the interrelatedness of samples and to extended regions of strong LD so only unrelated HapMap samples were included and a set of known long LD chromosome

regions were removed prior to analysis (Price, Patterson et al. 2006). A subset of SNPs that were also typed in our study samples was then extracted and the Eigenstrat SmartPCA program (Price, Patterson et al. 2006) was used to identify SNPs that best describe the differences between ethnic groups. These SNPs are then used to identify the ethnic origin of our samples. Those who do not cluster with the European HapMap populations and the rest of the sample set can be excluded prior to analysis.

The results of the supervised PCA (figure 7) show that no ethnic stratification was found in the ORCADES study and the samples correspond as expected to European descent.

Figure 7. Supervised PCA showing ethnic origin of the ORCADES population



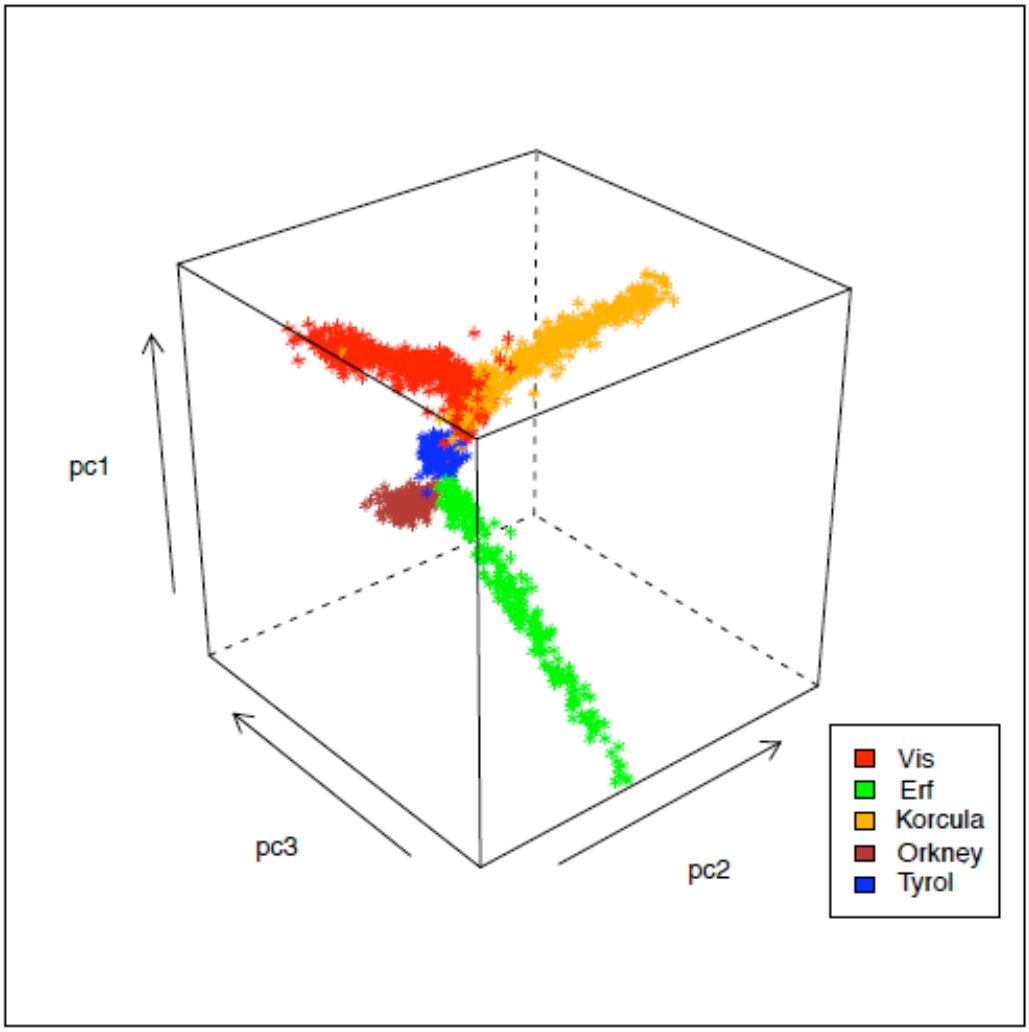
PCA can also be used to explore the more fine scale differences between our separate study populations. As with the HapMap analysis, closely related individuals may result in distortion of the major principle components and, since in this case the study populations are being used to define the PCs, the closely related individuals must be removed. To achieve this, genome-wide IBD estimates were calculated using PLINK

(Purcell, Neale et al. 2007). The average proportion of sites shared IBD between two individuals can then be used to estimate how closely related those samples are to one another, e.g. the expected proportion for full siblings is 0.5, half siblings 0.25 and first cousins 0.125.

To prune the close relationships from our populations we excluded one of each pair of samples with a proportion of IBD greater than 0.1. The remaining unrelated samples from each of the five populations were then combined. As with the HapMap supervised analysis regions of extended LD were excluded prior to running the PCA.

The results (figure 8) demonstrate that while all the study populations are European in ancestry there are still substantial genetic differences between populations. The first PC shows divisions between the northern European populations (ORCADES and ERF) and the southern European populations (VIS and KORČULA), the second PC shows separation between the two Dalmatian islands and the third PC shows some separation of the ORCADES and ERF populations. These differences make a pooled analysis unwise.

Figure 8. Unsupervised PCA showing 3-dimentional separation of the 5 study populations used in GWAS meta-analysis



4.3.2 Comparison of association methodologies

The inter-relatedness of the participants within each of our study populations will result in a large amount of phenotypic correlation between relatives due to the shared genetic background and heritability of the traits. This covariance would result in inflated evidence for association with a large number of markers which were segregating within a family but which have no actual influence on the traits in question, i.e. false positive results. To control for this effect we can use a relationship matrix to estimate the proportion of phenotypic variance which is due to shared genetic background. We can then use residual phenotypic values excluding this polygenic variance to test for association with an individual marker.

The majority of the studies being used here have collected large amounts pedigree information on their respective participants meaning that, along with genotypic information, we have two alternative sources of relationship information.

Kinship matrices based on known pedigree information are relatively simple to calculate and have been in use for a long time both in human and animal genetic research. The kinship score between two individuals represents the expected proportion of the genome, which will be shared identical-by-descent (IBD) between the related individuals. This is an average figure for a given pedigree relationship and the true level of sharing can be considerably larger or smaller due to the random nature of crossing over during meiosis.

With high density SNP data covering essentially the whole genome it is possible to estimate the level of genetic relatedness between two individuals based on their identity-by-state (IBS) sharing. By calculating the observed allele frequencies within our populations at each locus we can estimate the average proportion of all loci that we would expect to be shared between two unrelated individuals within that population. Shared ancestry will result in excess allele sharing and the degree of sharing can be used to estimate how closely the two individuals are related. Kinship matrices generated using this method can therefore give a more accurate estimate of the proportion of phenotypic variance due to polygenic effects. As an example of the differences in kinship that are revealed using this genomic estimator, sibling pairs on average share 50% of their genomes IBD, however the realized sharing for siblings in the ORCADES study varied between approximately 38% and 62%.

As a baseline for comparison of the alternative correction methods an initial analysis of FG, with age sex and BMI as covariates, was performed. The FG phenotype was rank-transformed prior to analysis to give an appropriate normal distribution. The results for this uncorrected analysis are shown in figure 9. The λ inflation of these results is approximately 1.09. Using genomic control to correct for this inflation gives the results shown in figure 10. As the same inflation factor is applied to all tests the relative ranking of SNPs is unchanged by GC and only the magnitude of the P-values is altered.

Figure 9. GWAS for ORCADES FG levels with age sex and BMI as covariates

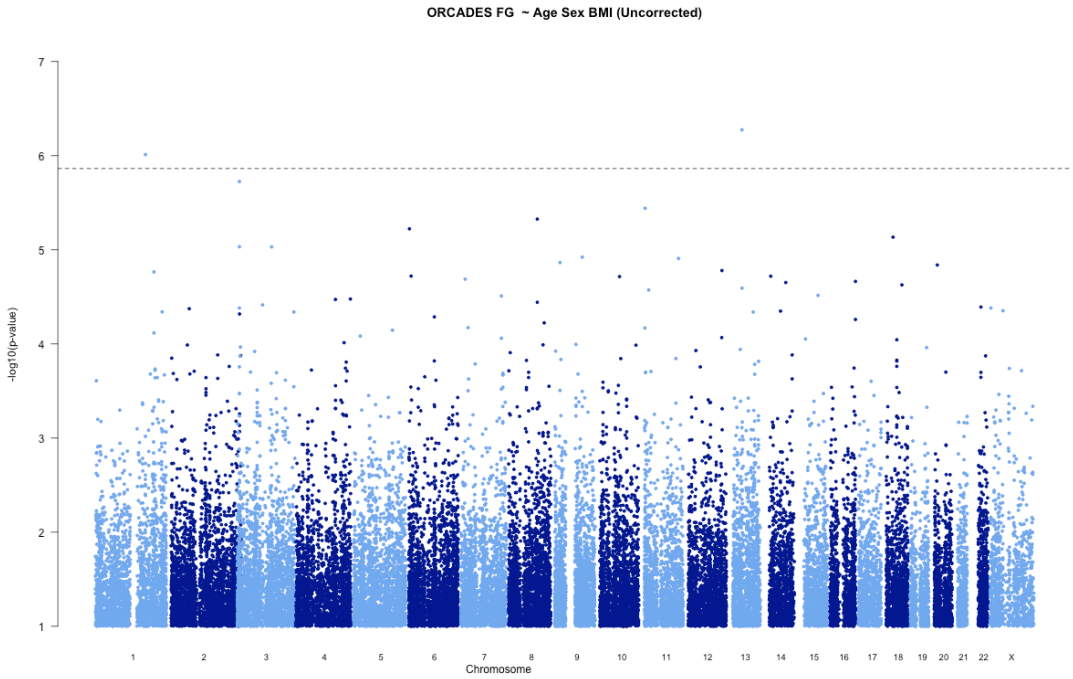


Figure 10. GWAS for ORCADES FG levels after GC correction with λ=1.09

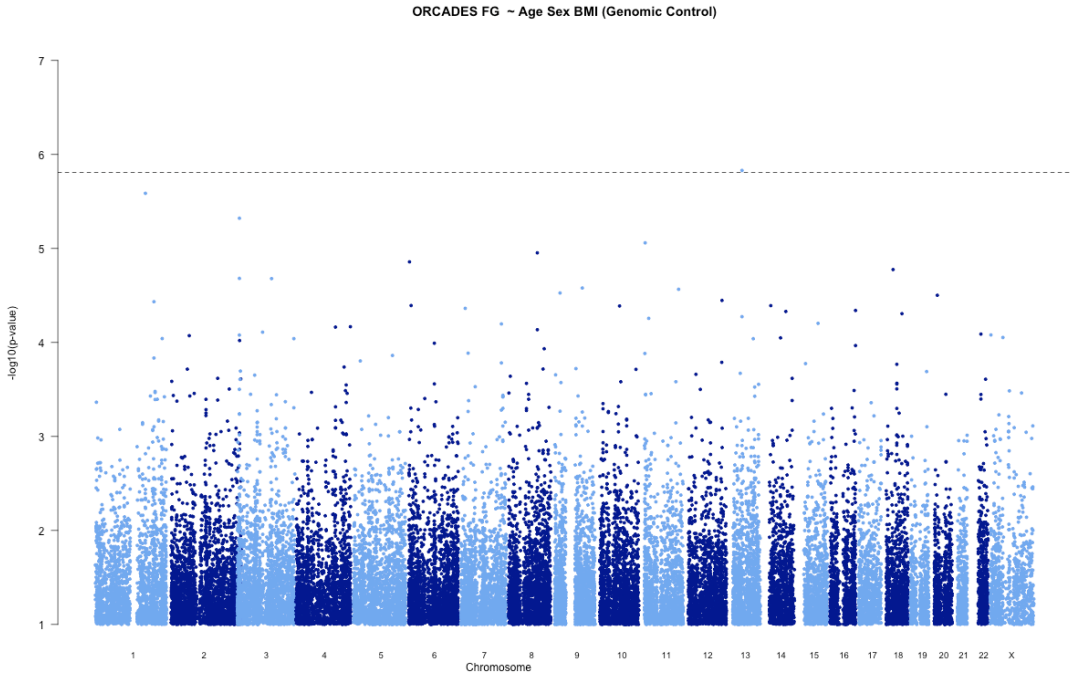
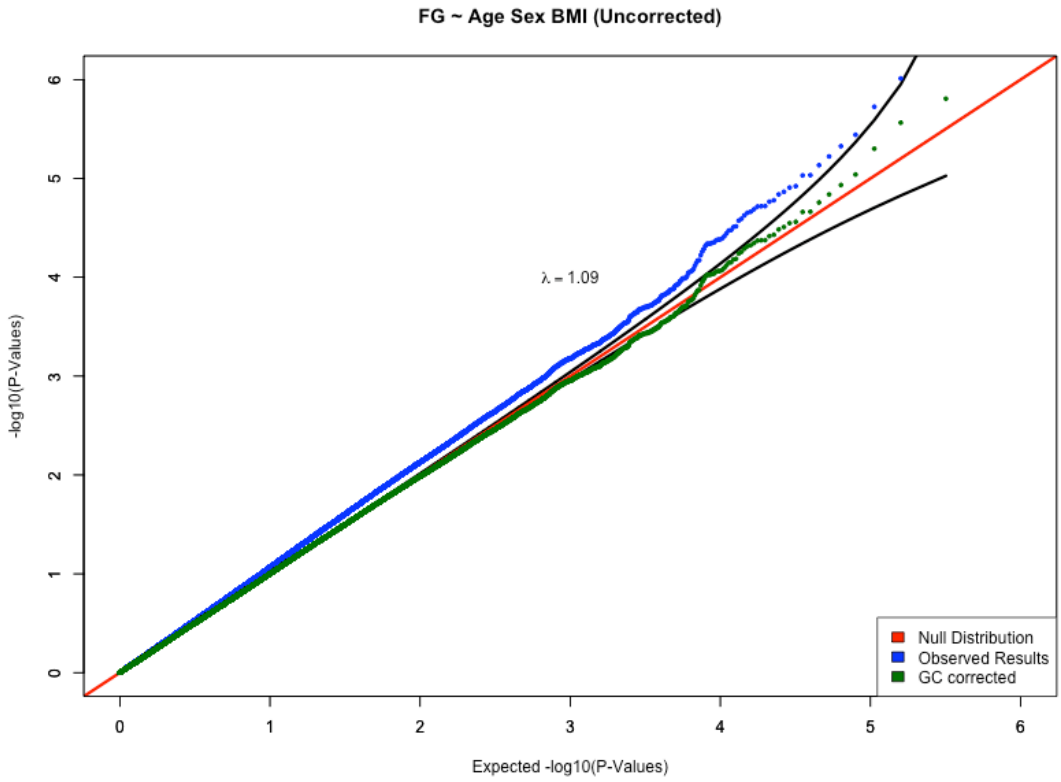


Figure 11 shows the QQ plots for these uncorrected and GC corrected GWAS results. The λ value of 1.09 reflects a considerable level of inflation across the entire range of test statistics.

Figure 11. QQ plot of Uncorrected and GC corrected P-values for ORCADES FG levels



The results obtained under the GRAMMAR methodology are shown in figure 12. The results profile changed considerably under this model with the most significant SNP result from the GC analysis greatly reduced while other markers have become considerably more significant. The QQ plot of the GRAMMAR analysis (figure 13) shows that a lower level of inflation is still present, with a λ value of 1.05.

Figure 12. GWAS for residual FG levels after correction for pedigree based kinship.

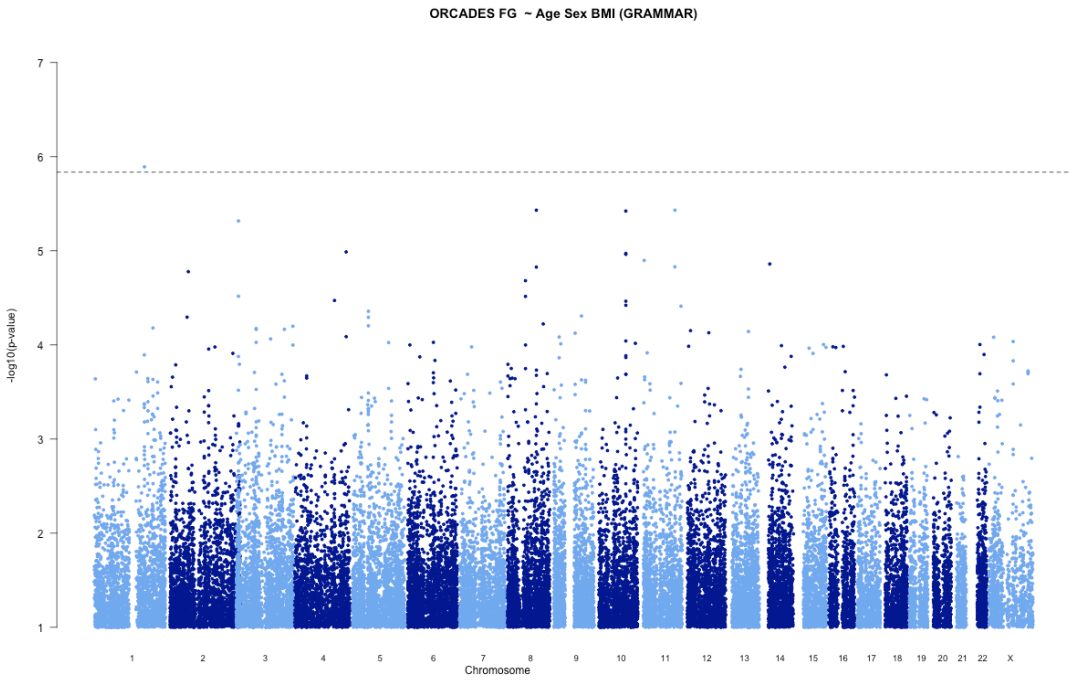
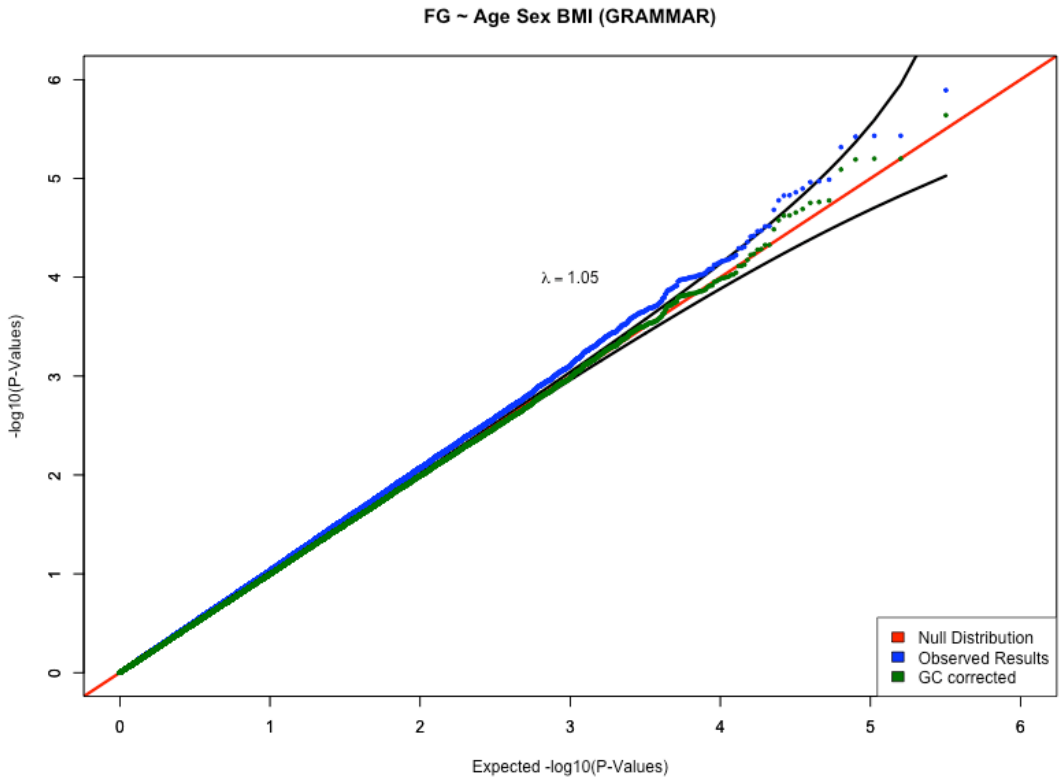


Figure 13. QQ plot of ORCADES FG results under the GRAMMAR method



The results obtained using the FASTA method are shown in figure 14. The resulting QQ distribution (figure 15) from this analysis showed no evidence of overall inflation with a λ value of 0.99.

Figure 14. GWAS results for ORCADES FG using genomic kinship correction under the FASTA methodology

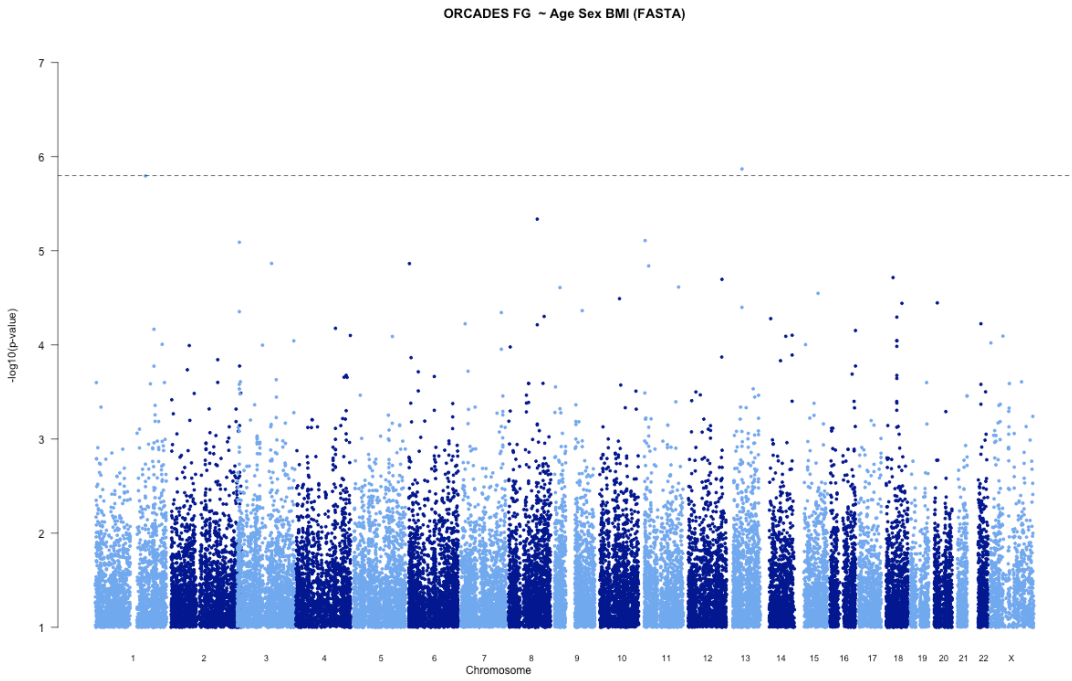
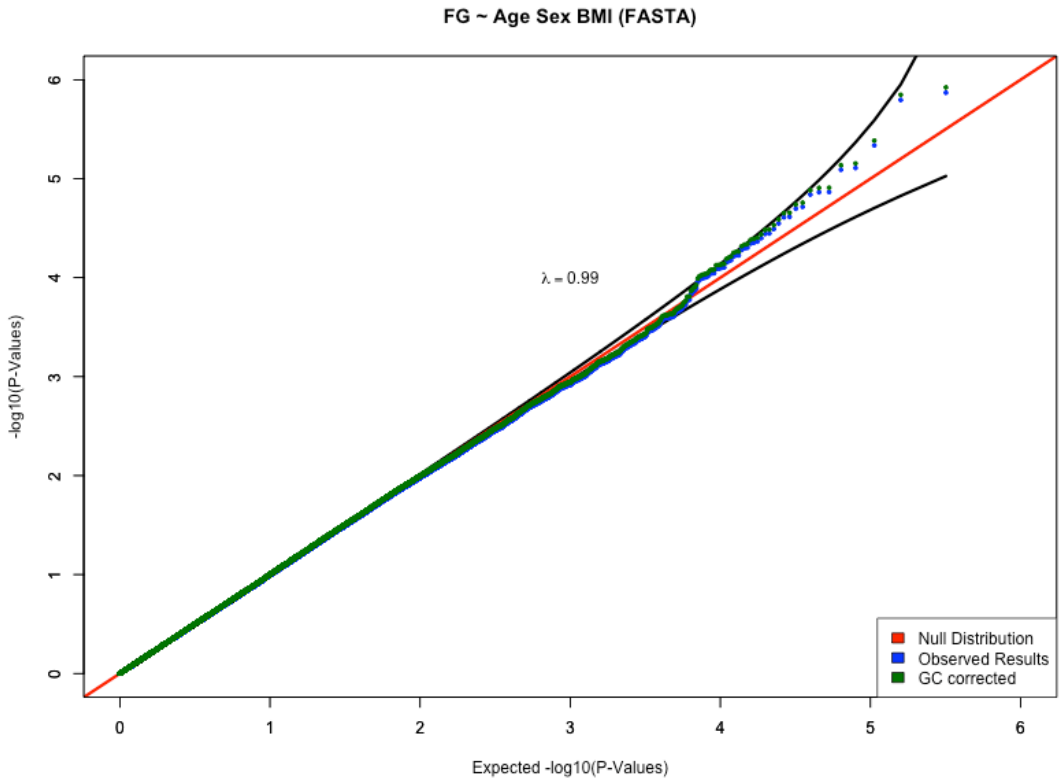


Figure 15. QQ plot of ORCADES FG results using FASTA

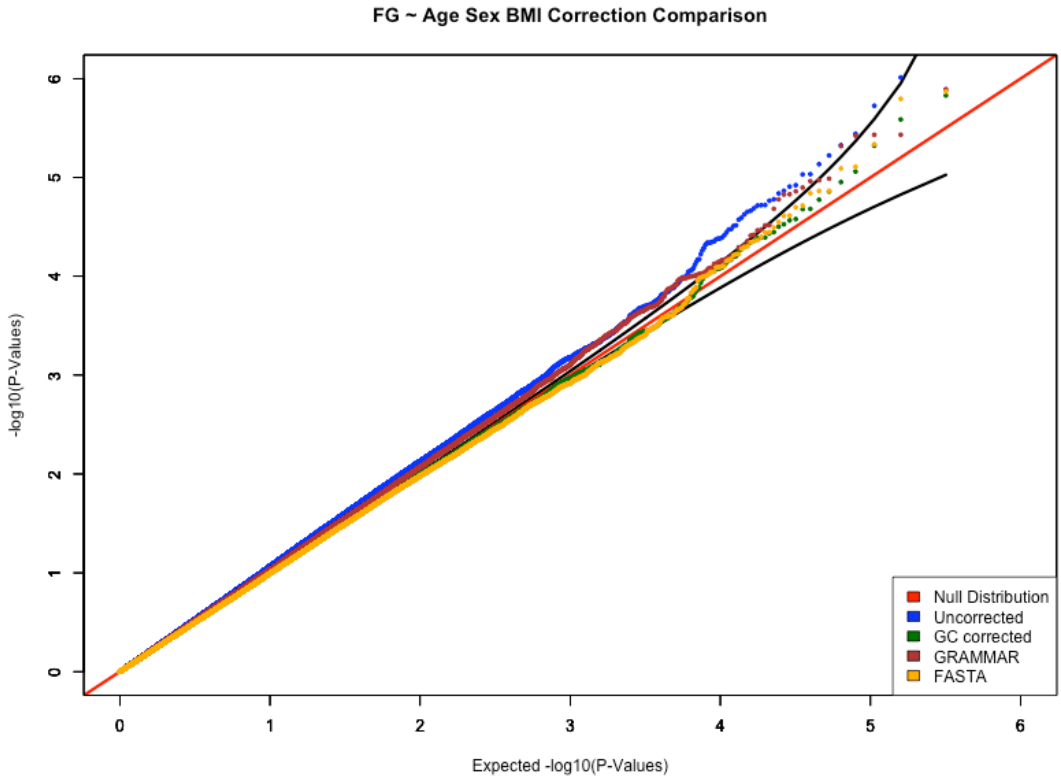


Finally, figure 16 compares the QQ plots of all the investigated correction methods. The relative λ values of 1.05 and 0.99 obtained using the GRAMMAR and FASTA methods respectively suggests that the genomic kinship correction step implemented in FASTA was more successful in breaking down the correlation between overall relatedness and phenotypic covariance, and was therefore less likely to produce spurious results or overly inflated effect estimates based on the population structure present in the ORCADES study.

Additionally, the use of the FASTA methodology, which does not require pedigree data, allowed for the use of standardized analysis methods when meta-analyzing

results including the KORČULA study for which no pedigree information was available.

Figure 16. QQ plot comparing ORCADES FG results obtained using GC, GRAMMAR or FASTA correction methods



4.4 Meta-analysis Results

Having chosen the FASTA method of analysis based on the ORCADES population, the diabetes-related phenotypes were tested for association in several further population isolates and meta-analysis was performed. The populations used are

described fully in the materials and methods chapter. In total there were 3 measured biochemical phenotypes, fasting glucose (FG), fasting insulin (FI) and HbA_{1C}. A further 2 phenotypes were derived from the ratio and product of FG and FI to give measurements of insulin resistance (HOMA-IR) and beta cell function (HOMA-B). HOMA-IR was calculated as (FG mmol/l * FI microunits/l)/22.5. HOMA-B was calculated as (FI microunits/l * 20)/(FG mmol/l - 3.5) (Matthews, Hosker et al. 1985).

Not all of the study populations had measured all 3 of the biochemical phenotypes so the number of samples, and the related power for association, for each analysis differs substantially. FG was available in the ORCADES, VIS, ERF, MICROS and KORČULA populations giving a combined sample of 4417 phenotyped individuals. FI and the derived HOMA traits were available in the ORCADES, VIS, and ERF populations totaling 2432 samples. HbA_{1C} was available in the ORCADES, VIS and KORČULA populations with a total sample size of 2402.

ORCADES biochemical measurements were from gel-separated serum or whole blood samples taken between 8am and 9:30am after fasting from 10pm the previous night. The samples were taken to the biochemical lab within 2 hours of being drawn and were run through a Beckman Coulter UniCel DxC600i analyzer. The coefficient of variation for the FG assay was 3.8% and the sensitivity of the FI assay was 0.03 uIU / ml. The HbA_{1C} measurements were taken from whole blood and were aligned

to the International Federation of Clinical Chemistry (IFCC) standards (Manley, John et al. 2004).

VIS measurements were from gel-separated serum or whole blood drawn between 8am and 9am after overnight fast. The samples were allowed to clot for 30 minutes before being centrifuged, aliquoted and stored at -70 degrees Celsius. The FG measures were obtained using UV hexokinase photometry and FI measurements were obtained using the ECL1A electrochemiluminescence immunoassay from Cobas. The sensitivity of this assay is 0.02 uIU / ml. The HbA_{1C} measurement were obtained by immunochemistry electrophoresis according to Diabetes Control and Complications Trial (DCCT) standards (1986).

The ERF samples were taken after overnight fast, serum separated and immediately frozen. The FG measurements were obtained from hexokinase assay using a Synchron LX20 and the FI measurements were from a Biosource INS-Irma kit. This assay has a reported sensitivity of 1 uIU / ml.

The MICROS samples were drawn from blood plasma after overnight fast between the hours of 7am and 9:30am and were taken to the biochemical lab within 3 hours. The FG measurements for this study were obtained using a hexokinase assay on a Dimension RxL.

The KORČULA FG measurements were obtained from gel separated serum or whole blood samples taken after overnight fast and were assayed by hexokinase photometry. An HbA_{1C} immunochemical electrophoresis assay was used and the results were aligned to the DCCT guidelines.

As HbA_{1C} levels were measured and reported according to IFCC guidelines in the ORCADES study and the measurements for the VIS and KORČULA studies were DCCT aligned, the ORCADES values were adjusted to the DCCT scale using the following conversion factor:

DCCT HbA_{1C} = (IFCC HbA_{1C} * 0.9148) + 2.152

Individuals with type 1 diabetes are always treated with insulin injections. T2D patients are usually treated initially non-pharmacologically through dietary advice. If however, dietary changes fail to control the individual's hyperglycaemia then oral glucose lowering drugs will be prescribed and in a small proportion of T2D cases insulin injections may become necessary.

Pharmacological interventions will results in altered levels of the glycaemic traits to those that would be predicted from their genetic and other environmental factors. In this case we are attempting to identify and quantify the genetic factors so not controlling for the medication effects may confound our ability to detect the genes in question.

There are several commonly used methods to control for medication effects in quantitative trait analysis as described in the section 2.6. In the case of diabetes the proportion of medicated individuals in our studies was between 2.1% and 9.2% and ultimately the simplest method was to exclude from the main analysis anyone who was taking oral medication or insulin injections.

Descriptive statistics for all the glycaemic phenotypes and relevant covariates are shown in table 5. The distribution of measured traits and covariates varied considerably between populations. All studies had a higher proportion of females than males. The MICROS population had the lowest mean age of 45 years and the VIS population was oldest with a mean age of 55 years. Mean BMI ranged from 25.6 kg/m^2 in MICROS to 27.9 kg/m^2 in KORČULA. The usage of diabetic medications and hyperglycaemia was highest in the two Croatian island populations with over 9% of the VIS study sample taking some form of glucose lowering medication. The ORCADES population had the lowest levels of glucose lowering medication with only 2.1%. The prevalence of diagnosed diabetes in Scotland overall is estimated to be 3.9%(McKnight, Morris et al. 2008).

Table 5. Descriptive statistics for the 5 Genome Wide Association studies.

	ERF	Orkney	Tyrol	Vis	KORČULA
N	918	719	1097	795	888
%Female	61.4	53.5	56.7	58.7	63.9
Age (years)	53.5 (15.2)	53.6 (15.7)	45.3 (16.1)	56.6 (15.4)	56.2 (13.9)
BMI (kg/m^2)	26.9 (4.7)	27.8 (4.9)	25.6 (4.83)	27.3 (4.25)	27.9 (4.15)
Fasting Glucose (mmol/l)	4.7 (1.1)	5.4 (0.98)	4.7 (1.08)	5.7 (1.49)	5.8 (1.55)
Insulin (mU/l)	12.6 (6.7)	6.7 (4.9)	na	9.4 (19.4)	na
HOMA-IR	2.6 (1.76)	1.5 (1.09)	na	2.4 (7.06)	na
HOMA-B	640 (4069)	71 (40)	na	88 (342)	na
HbA_{1C} (%)	na	3.7 (0.8)	na	5.5 (0.9)	5.8 (0.8)
Diabetes Medication(%)	4.9	2.1	3.6	9.2	6.3

4.4.1 FG – ORCADES, Vis, KORČULA, ERF, MICROS

FG measurements are the most widely used diagnostic test for T2D. Levels over 7 mmol/l on more than one occasion suggest a serious breakdown in glucose homeostasis and are used to diagnose diabetes.

A more accurate measure of glucose response can be obtained by an Oral Glucose Tolerance Test (OGTT) which tracks an individual's plasma glucose levels for 2hrs after a glucose challenge. Although desirable for the greater amount of information obtained this test is more time consuming than a single blood test. It also involves either multiple venepunctures or cannulation.

As a commonly used diagnostic test, and because of the relative ease of measurement, FG is the more widely studied in a genetic context than the other T2D related phenotypes. In recent years many independent genome-wide association studies have been conducted on FG levels and large-scale meta-analyses of these studies have been performed(Prokopenko, Langenberg et al. 2009; Dupuis, Langenberg et al. 2010). These meta-analyses have provided a number of well-replicated associations, which are now the subject of more detailed molecular study.

The results from our own meta-analysis are interesting both in the context of replicating well known associations, which can be considered positive controls for the quality of our study, and by providing novel associations not identified in the larger meta-analysis studies.

The first instinct when observing an association in a study of 4000 individuals, which does not appear in a larger analysis totaling over 100,000 samples, would be that the result is a false positive appearing by chance in our data and this may indeed be the case. However, the studies used here were all collected from genetically isolated populations in the hopes that allele frequency changes within these populations may allow detection of variants that are rarer in the larger European population. They could also potentially have a reduced number of genetic influences some of which may have been lost due to genetic drift(Sheffield, Stone et al. 1998). Coupled with the arguably more uniform environment found within these small populations this

should lead to stronger relative effect sizes for the polymorphisms which do still exist(Shifman and Darvasi 2001; Heutink and Oostra 2002).

For these reasons two sets of results from our meta-analysis association scans are of interest. Firstly, the highest scoring association results overall, and secondly the association results obtained for well-known, replicated FG polymorphisms. As might be expected in some cases these two lists will overlap.

Figure 17. GWAS meta-analysis results for FG in 5 populations with age sex and BMI as covariates

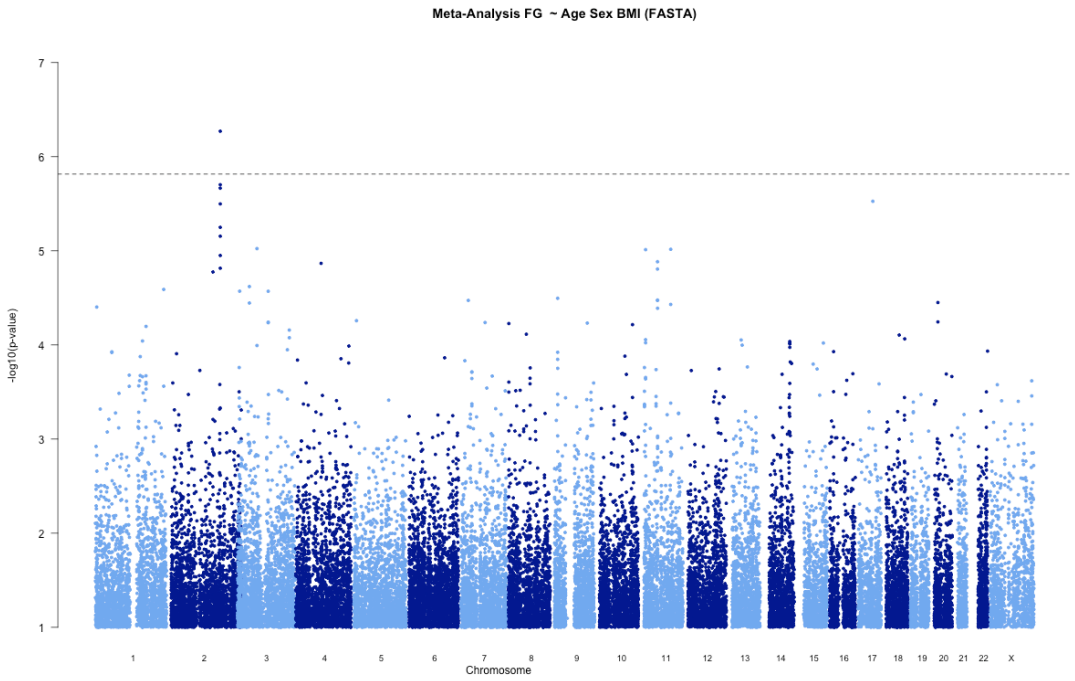


Figure 18. QQ plot of Fasting Glucose meta-analysis results

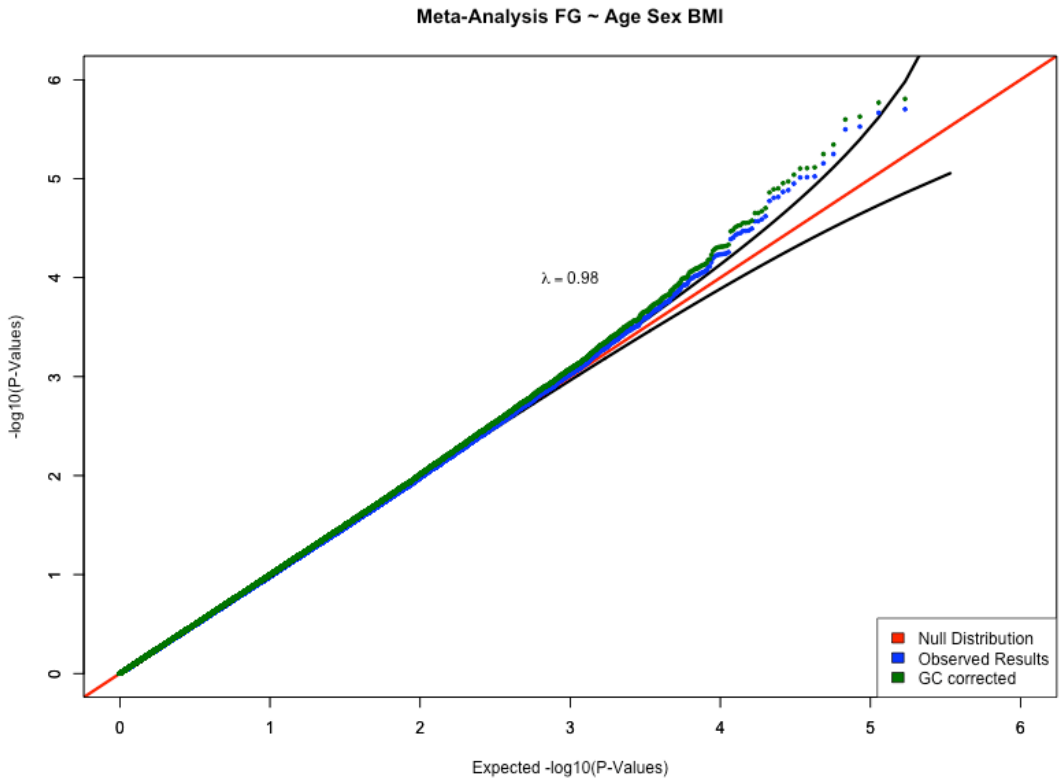


Table 6 shows the highest scoring associations for FG using a model including age, sex and BMI as covariates. The cut-off point for the table was a p-value less than $1x10^{-5}$. The strongest evidence for association is with SNP rs560887 on chromosome 2. A further five of the ten most highly associated SNPs are in the same chromosomal region. This region contains the gene Glucose-6-phosphatase 2 (*G6PC2*) for which association has been previously reported with both FG and HbA_{1C} (Chen, Erdos et al. 2008; Pare, Chasman et al. 2008). Two of the six associated markers also show strong evidence for association in the female-only analysis but not the male specific analysis. This could represent a genuine differential sex effect for these polymorphisms, but alternatively could reflect

differences in power arising from the gender skew present in the study populations, which results in considerably larger numbers for the female specific analyses.

The marker rs13078632 on chromosome 3 also shows association with FG ($p=9.48x10^{-6}$). This marker is located in the 3-prime un-translated region of the Kelch Repeat and BTB domain-containing protein 8 (*KBTBD8*), which is involved in T-cell activation.

Two markers on chromosome 11 show association. One of these markers, rs1387153, has previously been identified in large-scale analysis of FG and T2D. This variant is located near the Melatonin Receptor 1B (*MTNR1B*). The other chromosome 11 association (rs17372123) falls within the gene encoding Haemoglobin subunit gamma-2 (*HBG2*). The last association found in analysis of all individuals is rs2907666 that lies approximately 150 kb upstream of a cluster of genes including *TOM1L1*, *COX11* and *STXBP4*, none of which have an obvious role in glycaemic control.

The sex-specific analysis was performed using age and BMI as covariates. There is some overlap between genes that show up in the overall analysis and those that are most significant in the sex-specific analysis but there are also some markers that show association only in one gender. This could represent a genetic effect that is mediated in someway by transcriptional or hormonal differences between the genders. It must also be considered that with the reduced sample sizes for the sex

specific analysis power to detect true associations is reduced and the chance of false positive results increases.

In addition to the associations with *MTNR1B* and *G6PC2* already found in the non-stratified analysis the female-only meta-analysis shows associations on chromosome 3 with the Glutamate Receptor 7 (*GRM*) and a separate signal within an intron of the Chemokine XC Receptor (*XCR1*). There are also associations on chromosome 16 with Ataxin-2-binding Protein (*A2BP1*) and on chromosome 22 with a particularly gene-rich region containing several uncharacterized protein-coding genes.

The analysis of only the male subjects yielded just one association with a p-value less than 1×10^{-5} on chromosome 8. This marker does not fall within any gene but is downstream of several genes involved in signal trafficking.

Table 6. Top association results for fasting glucose in 5-population meta-analysis.

Gene	SNP	Allele	Freq	Chr	Position (mb)	N	Effect	S.E.	p-value
G6PC2	rs3931	G	0.28	2	169.5	3940	-0.08	0.02	2.16x10-6
G6PC2	rs560887	G	0.71	2	169.6	3975	0.08	0.02	5.36x10-7
G6PC2	rs563694	C	0.36	2	169.6	3966	-0.06	0.01	5.64x10-6
G6PC2	rs503931	C	0.52	2	169.6	3972	0.05	0.01	7.00x10-6
G6PC2	rs2685814	G	0.53	2	169.6	3952	0.05	0.01	3.18x10-6
G6PC2	rs853778	G	0.52	2	169.6	3975	0.06	0.01	1.99x10-6
KBTBD8	rs13078632	G	0.14	3	67.1	3972	-0.12	0.03	9.48x10-6
MTNR1B	rs17372123	C	0.33	11	5.5	3961	0.07	0.02	9.72x10-6
HBG2	rs1387153	G	0.73	11	92.3	3974	-0.08	0.02	9.65x10-6
TOM1L1	rs2907666	G	0.83	17	50.2	3967	-0.11	0.02	2.98x10-6

Other large-scale analyses have identified a large catalogue of FG loci during the period of this phd project. Examining the most highly significant SNP markers from these studies I identified 16 loci for which we had genotyped the reported marker itself or nearby proxy SNPs in strong LD. The meta-analysis results for these 16 loci are shown in table 7. As previously mentioned, association with the *MTNR1B* and *G6PC2* genes showed strongly in our data. There were also weaker association signals with the Glucokinase gene (*GCK*) and its receptor (*GCKR*). Although these

genes have previously recognized, and entirely logical, roles in FG control they would not have been identified in this analysis of nearly 4000 individuals.

Table 7. 5-population meta-analysis association results for previously identified fasting glucose markers.

Gene	Proxy	Distance	R2	Allele	Freq	N	Effect	S.E.	p-value
ADCY5	rs2877716	28673	0.82	G	0.76	3972	0.03	0.02	0.16
ADRA2A	rs10787315	9367	0.82	G	0.92	3945	0.10	0.04	4.53x10-3
C2CD4B	rs12440695	1194	1.00	G	0.37	3975	-0.01	0.01	0.69
DGKB-TMEM195	rs2191348	54	1.00	C	0.45	3969	-0.03	0.01	0.03
DGKB-TMEM195	rs10244051	476	1.00	C	0.55	3976	0.02	0.01	0.04
FADS1	rs174546	1648	1.00	G	0.67	3976	0.02	0.02	0.20
FADS1	rs174556	9157	0.80	G	0.72	3975	0.01	0.02	0.45
FADS1	rs102275	13675	1.00	G	0.34	3975	-0.02	0.02	0.13
FADS1	rs174537	18798	1.00	C	0.68	3949	0.02	0.02	0.15
FADS1	rs1535	26494	1.00	G	0.33	3974	-0.02	0.02	0.12
G6PC2	rs560887	0	1.00	G	0.71	3975	0.08	0.02	5.36x10-7
G6PC2	rs563694	0	1.00	C	0.36	3966	-0.06	0.01	5.64x10-7
GCK	rs4607517	0	1.00	G	0.81	3967	-0.07	0.02	7.76x10-4
GCKR	rs780094	0	1.00	G	0.61	3969	0.03	0.01	0.04
GCKR	rs1260333	7387	0.87	G	0.55	2185	0.01	0.02	0.51

GCKR	rs1260326	10297	0.93	G	0.59	3245	0.03	0.01	0.07
GLIS3	rs7041847	1584	0.94	G	0.48	3975	0.00	0.01	0.81
GLIS3	rs10814916	4100	0.81	C	0.53	3974	0.00	0.01	0.79
MADD	rs11039149	59645	1.00	G	0.31	3972	-0.04	0.02	0.03
MTNR1B	rs1387153	0	1.00	G	0.73	3974	-0.08	0.02	9.65x10-6
PROX1	rs340874	0	1.00	G	0.53	3973	0.02	0.01	0.11
SLC2A2	rs5400	14779	1.00	G	0.87	3957	0.08	0.03	2.69x10-3
SLC2A2	rs11919048	143356	0.87	G	0.86	3972	0.05	0.03	0.09
SLC30A8	rs13266634	950	0.96	G	0.71	3972	0.03	0.02	0.06
TCF7L2	rs7901695	1953	0.88	G	0.31	3241	0.05	0.02	5.29x10-3
TCF7L2	rs7903146	2308	0.92	G	0.71	3975	-0.05	0.02	1.13x10-3
ZMAT4	rs2722425	0	1.00	G	0.89	3976	-0.03	0.03	0.38

4.4.2 FI – ORCADES, VIS, ERF

The FI analysis was conducted in a similar way to the FG analysis. Individuals taking diabetic medications were excluded, the highly skewed trait was rank-transformed to normality, variance attributable to polygenic effects was corrected for and association analysis was performed for each study population using the FASTA test.

In the meta-analysis, shown in figure 19, no individual SNP reached a genome-wide significant level of association and it can be seen from the QQ plot (figure 20) that the distribution of results showed no inflation when compared to the null.

Figure 19. GWAS meta-analysis results for FI levels in ORCADES, VIS and ERF populations with age sex and BMI as covariates

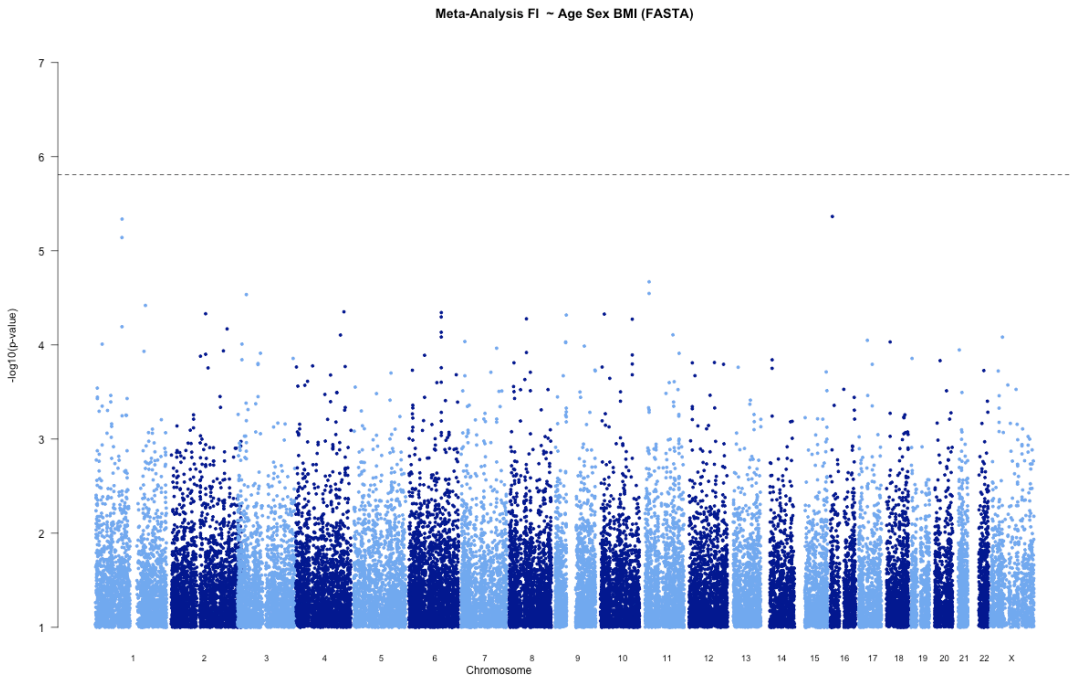
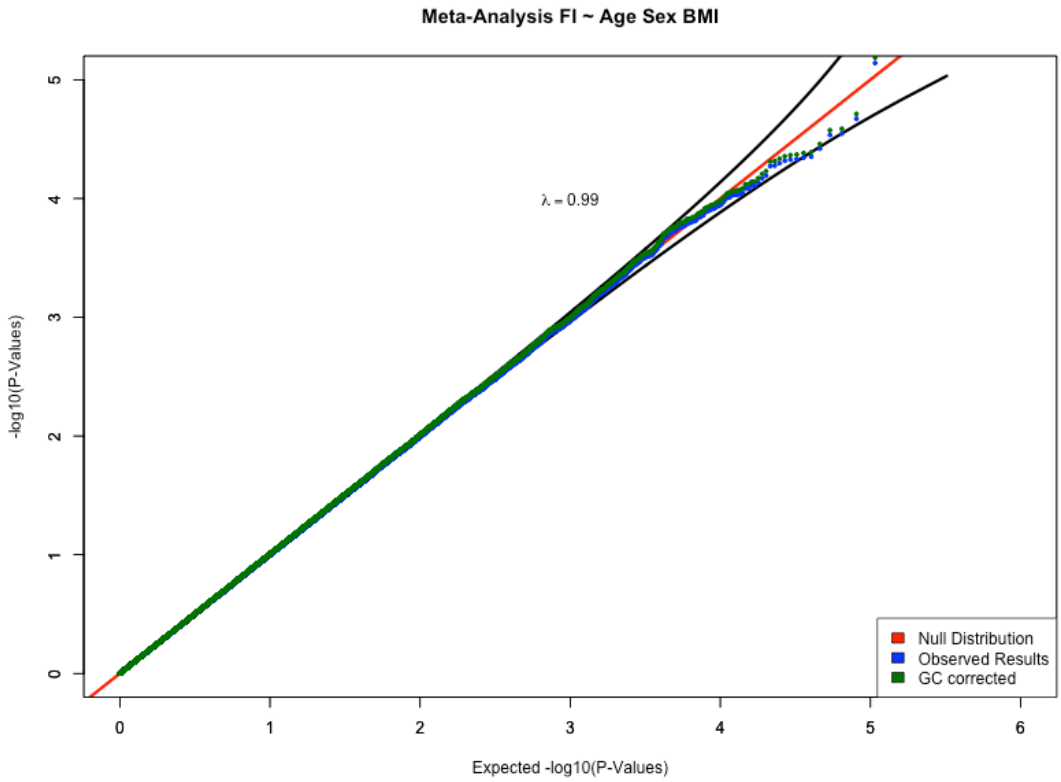


Figure 20. QQ plot of Fasting Insulin meta-analysis results



The strongest association result seen for FI is with SNP rs1936002 (table 8), which falls within an intron of the serine/threonine-protein kinase *DCLK1*. The other FI association found in the analysis of all individuals is rs3785690, which resides within an intron of the Heparan Sulfate Glucosamine 3-O-sulfotransferase 3A1 (*HS3ST3A1*).

In the sex-specific analysis of males no marker showed evidence for association beyond the $p < 1x10^{-5}$ threshold. In the female-only analysis there were 3 associated markers, one in the TBC1 domain family member 5 gene (*TBC1D5*) on chromosome

3, one downstream of the coiled-coil domain containing protein 60 (*CCDC60*) on chromosome 12 and one on chromosome 18 downstream of the *ROCK1* (Rho-associated, Coiled-coil containing protein kinase 1) gene which doesn't have any obvious role in insulin control.

Association results for previously reported FI SNPs and/or their closest available proxies are shown in table 9. My analysis shows nominal evidence of association at the Insulin Growth Factor (*IGF1*) locus but no evidence in support of the *GCKR* locus.

Table 8. FI meta-analysis association results

Gene	SNP	Allele	Freq	Chr	Position (mb)	N	Effect	S.E.	p-value
DCLK1	rs1936002	G	0.54	13	35.4	2115	0.15	0.03	6.38×10^{-6}
HS3ST3A1	rs3785690	G	0.84	17	13.4	2115	0.20	0.04	6.18×10^{-6}

Table 9. Meta-analysis results for previously reported FI loci. R^2 correlation with previously reported SNP based on HapMap CEU data.

Locus	Chr	SNP	Position (mb)	R^2	P-value
IGF1	12	rs2162679	101.4	0.92	0.006
IGF1	12	rs35766	101.4	0.92	0.009
GCKR	2	rs1260326	27.6	0.93	0.75
GCKR	2	rs1260333	27.7	0.87	0.49
GCKR	2	rs780094	27.6	1	0.82

4.4.3 HOMA-IR/B - ORCADES, VIS, ERF

HOMA-IR, calculated as a product of FI and FG levels, and HOMA-B, calculated as a ratio of FI to FG, were examined in the populations that had obtained both of these phenotypes. Both phenotypes have a strong skew and were transformed by natural logarithm prior to analysis.

The strongest association with HOMA-IR is the SNP rs3790451 (table 10), which lies within the Ecotropic Viral Integration site 5 (*EVI5*). There are several other protein coding genes within a few hundred kilobases of this site but none with well-explored functions.

The female-specific analysis of HOMA-IR showed 4 markers beyond the $p<1 \times 10^{-5}$ threshold level. There are two associations on chromosome 8 separated by 7 Mb. The first falls within a ribosomal gene *RPS15A* and the second is in the Hepatocyte Nuclear Factor 4-gamma (*HNF4G*) gene.

Two of the top 3 associations with HOMA-IR in the male-specific analysis cluster together on chromosome 1 in a gene-poor region. The other association found in the analysis is more promising, lying within the Insulin-like Growth Factor 1A precursor (*IGF1A*).

Figure 21. GWAS meta-analysis results for HOMA-IR in ORCADES, VIS and ERF populations with age sex and BMI as covariates

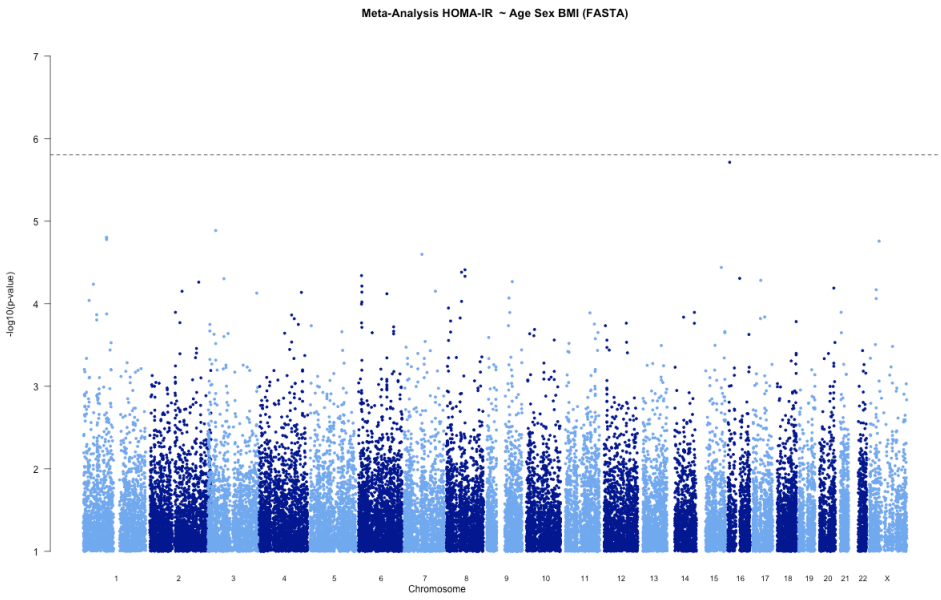


Figure 22. QQ plot of HOMA-IR meta-analysis results

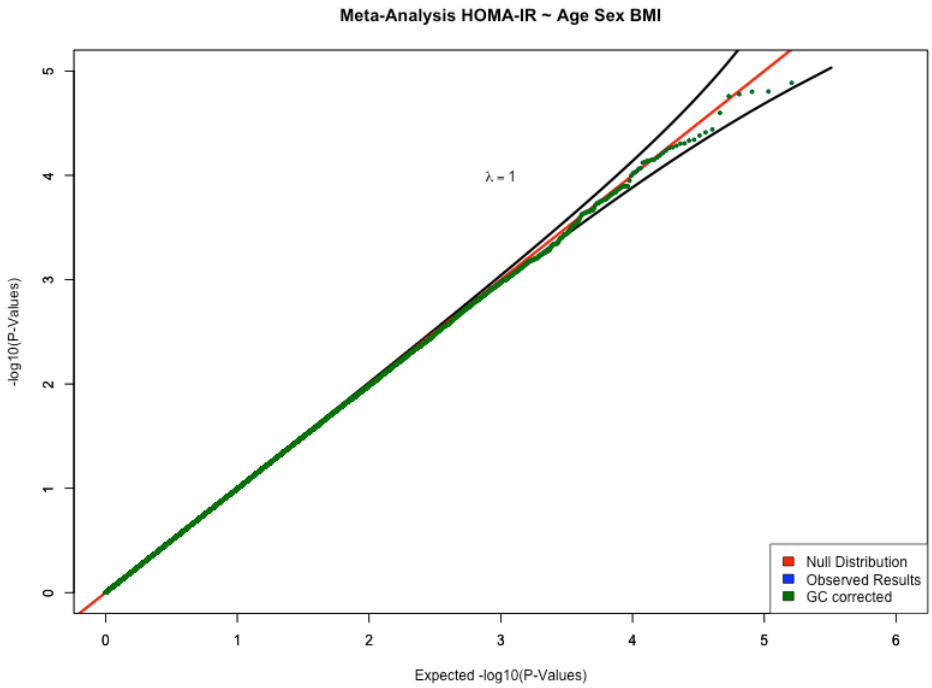


Figure 23. GWAS meta-analysis results for HOMA-B in ORCADES, VIS and ERF populations with age sex and BMI as covariates

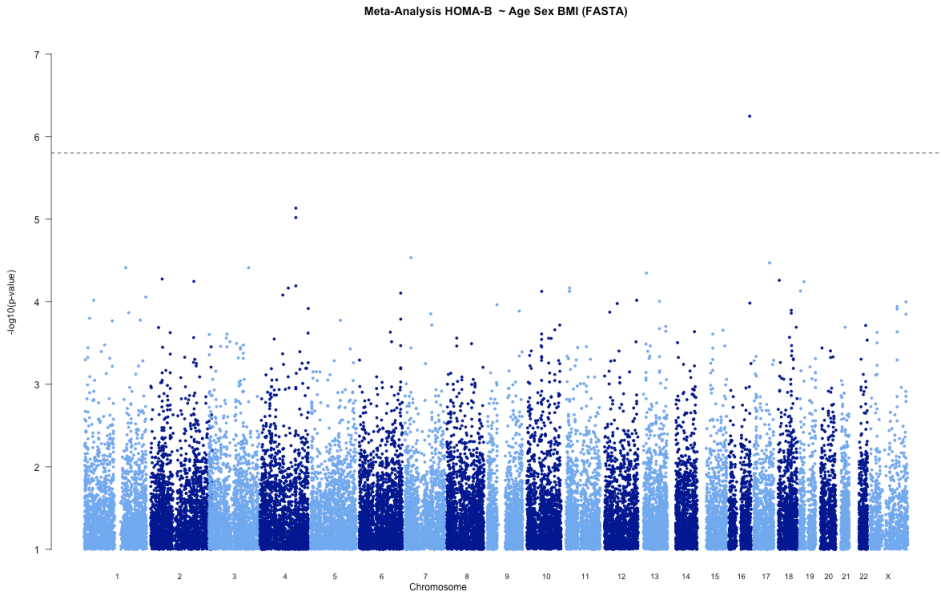


Figure 24. QQ plot of HOMA-B meta-analysis results

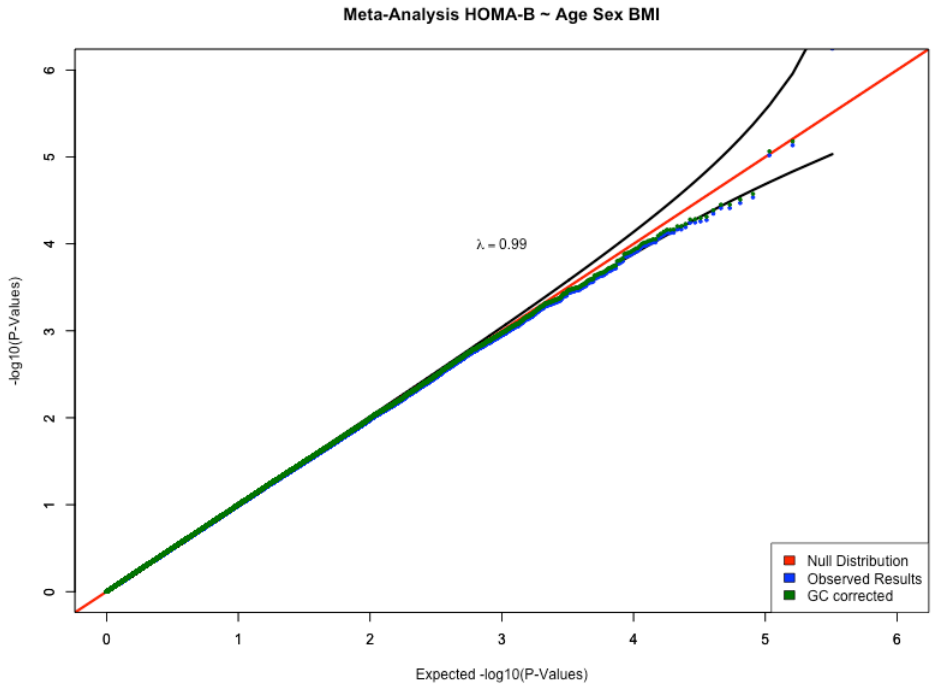


Table 10. HOMA-IR meta-analysis association results for ORCADES, VIS and ERF populations

Gene	SNP	Allele	Freq	Chr	Position (mb)	N	Effect	S.E.	p-value
EVI5	rs3790451	G	0.48	1	92.9	2093	-0.15	0.03	2.88×10^{-6}

The results obtained from HOMA-B analysis (table 11) showed an association on chromosome 8 with rs10504962 ($p=1.8 \times 10^{-6}$) upstream of the Syndecan (*SDC2*) locus. The second strongest association, rs2871491 ($p=4.1 \times 10^{-6}$), falls within the *SORB2* gene.

Table 11. HOMA-B meta-analysis association results for ORCADES, VIS and ERF populations

Gene	SNP	Allele	Freq	Chr	Position (mb)	N	Effect	S.E.	p-value
SORB2	rs2871491	G	0.82	4	186.9	2102	-0.2	0.04	4.12×10^{-6}
None	rs11132593	G	0.52	4	190.1	2096	-0.14	0.03	8.63×10^{-6}
SDC2	rs10504962	G	0.9	8	97.5	2083	0.26	0.05	1.79×10^{-6}
ETNK1	rs11046815	G	0.6	12	23.2	2101	0.15	0.03	9.05×10^{-6}

4.4.4 HbA_{1C} ORCADES, VIS, KORČULA

The association analysis of HbA_{1C} levels showed 3 hits above the Bonferroni correction threshold (figure 25). The two strongest association signals were with markers rs7903146 ($p = 1.48 \times 10^{-7}$) and rs12255372 ($p = 8.44 \times 10^{-7}$) in the *TCF7L2* locus (table 12). This result was of particular interest because, while these SNPs have previously been shown to associate strongly with T2D (Grant, Thorleifsson et al. 2006), they had not previously been shown to affect HbA_{1C} levels within the healthy non-diabetic population. The results of this analysis were published during the writing of this thesis and are attached in appendix A (Franklin, Aulchenko et al. 2010).

The association results from the individual studies are shown for rs7903146 in table 13. Effect estimates in this case are calculated based on untransformed HbA_{1C} levels and are therefore expressible in units of % HbA_{1C}. While the SNP does not approach genome-wide significance in any of the studies, it does show a consistent size and direction of effect.

The third most significant SNP (rs12366899, $p=1.14 \times 10^{-6}$) was only genotyped in the KORČULA study and, based on the very small number of samples is likely to be a false positive artifact.

Figure 25. GWAS meta-analysis results for HbA_{1C} in ORCADES, VIS and KORČULA populations with age sex and BMI as covariates

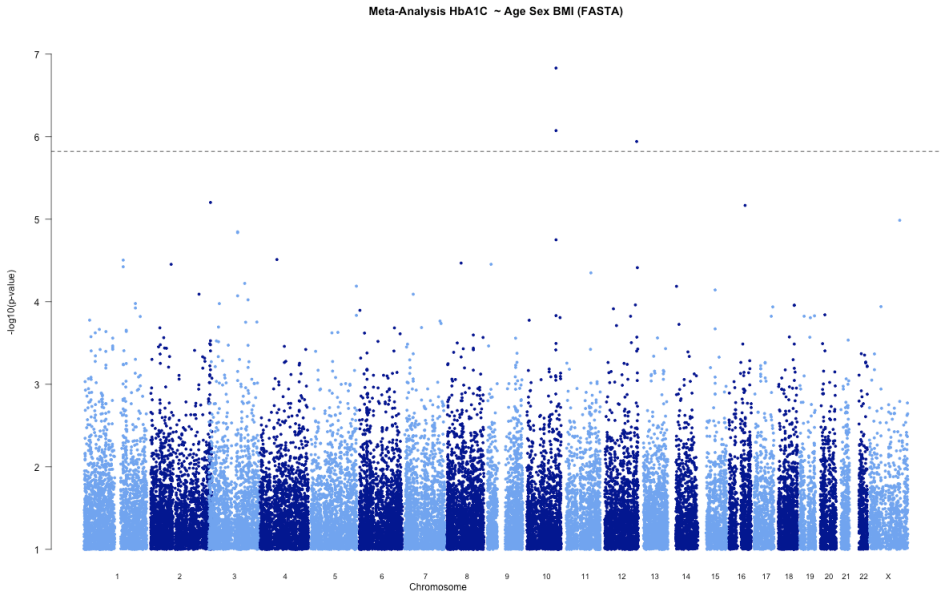


Figure 26. QQ plot of HbA_{1C} meta-analysis results

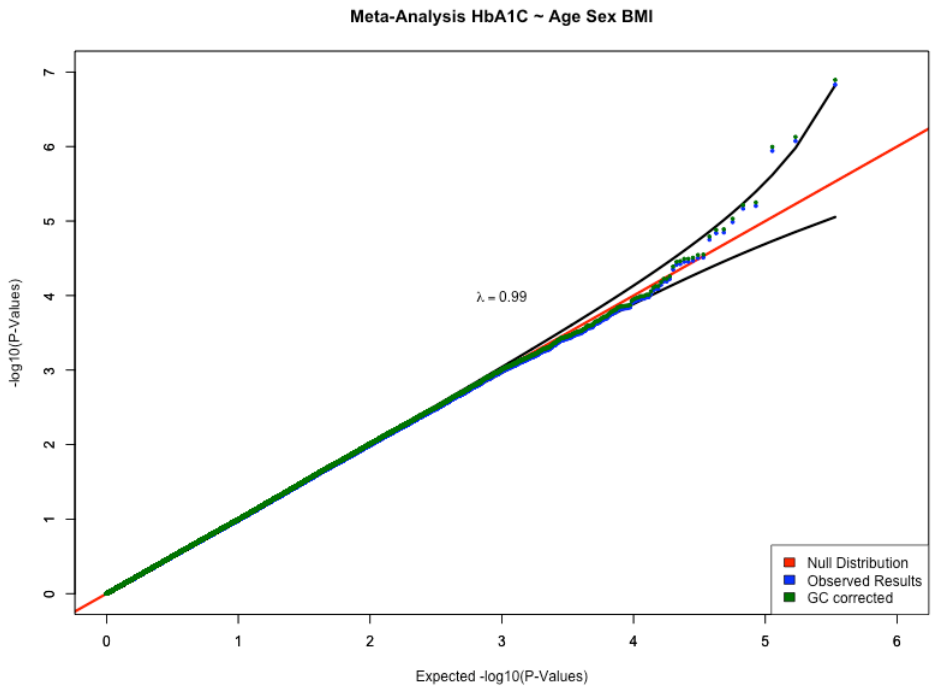


Table 12. HbA_{1C} meta-analysis association results for ORCADES, VIS and KORČULA populations

Gene	SNP	Allele	Frequency	Chr	Position (mb)	N	Effect	S.E	p-value
TCF7L2	rs7903146	G	0.72	10	114.7	1782	-0.20	0.04	$1.48x10^{-7}$
TCF7L2	rs12255372	C	0.73	10	114.8	1782	-0.19	0.04	$8.44x10^{-7}$
None	rs12366899	G	0.64	12	125.6	406	0.42	0.09	$1.14x10^{-6}$

Table 13. Association results for rs7903146 in individual populations.

(Statistics are reported for the G allele. Effect estimates and standard errors are calculated using untransformed HbA_{1C} values. P-values are calculated using rank transformed data.)

Population	N	Freq	Effect (%HbA_{1C})	S.E.	P-Value
Meta-Analysis	1782	0.72	-0.054	0.014	1.48×10^{-7}
ORCADES	664	0.74	-0.067	0.024	1.41×10^{-4}
VIS	707	0.71	-0.044	0.021	1.54×10^{-2}
KORČULA	411	0.71	-0.056	0.033	4.21×10^{-3}

4.5 Discussion

The analysis shown here demonstrates the benefits of meta-analysis. The individual studies had quite small sample sizes and were greatly underpowered to detect QTs with the very small effect sizes that we now know to expect. As a result the highest associations obtained from individual studies are quite likely to be false positives.

However, by pooling up to 5 of these small studies we were able to improve the power for detection and bring associations, which we know from larger concurrent studies to be genuine, to the top of our results tables.

Two of the most widely replicated FG genes, *G6PC2* ($p=5.36x10^{-7}$) and *MTNR1B* ($p=9.65x10^{-6}$), showed strong association in our meta-analysis and, in the absence of other research in the area, our own study would have identified and sought replication for these loci. Several more known loci showed nominally significant associations including *ADRA2A* ($p=4.53x10^{-3}$), *DGKB-TMEM195* (p=0.03), *GCK* ($p=7.76x10^{-4}$), *GCKR* (p=0.04), *MADD* (p=0.03), *SLC2A2* ($p=2.69x10^{-3}$) and *TCF7L2* ($p=5.29x10^{-3}$). While these loci may not have been taken forward for replication based on our own results they do still represent a form of positive control.

Perhaps the more interesting results are the known and well-replicated loci that do not show any evidence of association in our data. The *ADCY5*, *C2CD4B*, *FADS1*, *GLIS3*, *PROX1*, *SLC30A8* and *ZMAT* loci have all shown strong evidence for association in published FG GWAS analyses and yet the same reported SNP markers, or nearby proxy SNPs present in our data, do not show even nominal significance in our meta-analysis of 4,000 individuals.

This lack of replicating could indicate the limitations of our study. Even after the pooling of 5 studies our meta-analysis still has considerably lower power to detect, at genome-wide significance levels, variants with very small effect sizes than the very

large meta-analyses in which many of these loci were identified. However, even with a smaller sample our study should have substantial power to detect these variants at a nominal $p<0.05$ threshold.

Some alternate explanations would be that the nature of our isolated population study design results in different environmental modifiers than are found in cosmopolitan populations, and it is possible that some genetic variants will only have an effect on the phenotype in the presence of such modifiers.

In order to be confident of identifying QTs with small effect sizes, extremely large sample sizes are required. To this end we have joined a number of international consortia, which have been, and continue to conduct, very large meta-analyses of the traits discussed here. Depending on how commonly measured the traits are these analyses now consist of 40 to over 100 thousand samples, allowing the detection of variants with effect sizes as small as 0.25% of phenotypic variation in some cases.

The value in identifying genetic variants with such small individual effect sizes is somewhat different to the outcome of studying monogenic disorders. While each new discovery of this magnitude may not be revolutionary alone they do add to knowledge of the mechanisms of disease development and may lead to the investigation of pathways that may not otherwise have been explored.

Results obtained from association studies can be roughly divided into 3 groups:

1. Association with a gene or pathway that has a plausible role in studied trait or disease based on previously determined function.
2. Association with a gene or pathway, which does not have an obvious connection to the trait or disease being studied.
3. Association that falls far from any coding gene.

4.5.1 Genes with known function

The way in which some of the known glucose genes may be influencing glucose concentrations and related traits is quite straightforward. The *GCK* gene, for example, produces a type of hexokinase which is expressed primarily in the liver and which catalyses the first step in glucose metabolism (glycolysis). The enzyme activity of GCK is proportional to glucose concentration, increasing as glucose levels rise after a meal and bottoming out when the glucose levels are reduced to a normal level(Gloyn 2003). This function makes *GCK* potentially the most obvious candidate for association with fasting glucose levels.

Secondly the primary function of the GCKR protein is to regulate the activity of the GCK enzyme. It does this by binding the GCK in the cellular cytoplasm, preventing the binding of glucose and inactivating the enzyme. The GCKR then travels to the nucleus, reducing cellular GCK concentration. Rising cellular glucose concentration inhibits GCKR function therefore releasing GCK to perform it function(Beer, Tribble et al. 2009). Alterations to the function of GCKR which caused the protein to

become insensitive to glucose concentrations or which caused tighter binding of the GCK enzyme could easily result in hyperglycaemia and ultimately in the development of T2D.

The G6PC2 enzyme is primarily active in pancreatic cells. Whereas GCK is involved in the metabolism and storage of glucose and becomes active in higher than optimal glucose conditions, the G6PC2 enzyme is involved in the gluconeogenic and glycogenolytic pathways, which release stored glucose when blood glucose levels are too low(Hutton and O'Brien 2009). Again it is easy to see how a mutation which altered the normal regulation of this protein or which caused over expression of the coding gene could result in increased FG and over time cause a breakdown of glucose homeostasis leading to the development of T2D.

4.5.2 Genes with unknown function

MTNR1B is an excellent example of a gene which may not have been expected to show strong association with FG and may not have been chosen in candidate gene studies, prior to the large-scale GWAS analysis that first identified the link (Prokopenko, Langenberg et al. 2009). The investigators who identified the gene were however able to provide supporting biological evidence for the role, highlighting the known expression patterns of the *MTNR1B* gene, which demonstrate substantial expression in pancreatic islets (Ramracheya, Muller et al. 2008). This expression localization, coupled with evidence that melatonin inhibits the release of

insulin (Stumpf, Muhlbauer et al. 2008), demonstrates a reasonable explanation for the role of the melatonin receptor in regulation of blood glucose levels.

4.5.3 Associations in gene deserts

In some cases a statistically significant association may fall within a region that contains no known or predicted gene sequences. At first consideration this may seem less interesting. However as we are dealing with QTLs, which exert a small level of control over a continuous trait, it is important to consider regulatory elements that modify levels of gene expression. These can be positioned a great distance from the actual coding sequence which they act on(Hakim, John et al. 2009).

Many different methods have been attempted to identify promoter regions and regulatory elements from genome sequences. For example CpG islands can be searched for, this being a region in which a higher proportion of CG di-nucleotide pairs exist. This usually denotes a region that is actively transcribed(Fatemi, Pao et al. 2005).

An alternative method is comparison of sequences from related species to identify regions which are conserved across, for example higher primates or all mammalian species. A higher degree of sequence conservation suggests a region is important functionally(Hardison 2000), although rapidly evolving regions may of course also have important species-specific roles(Hardison, Krane et al. 1991).

4.5.4 Glycaemic phenotypes and T2D

As mentioned the study of continuously distributed glycaemic phenotypes within a normal healthy range is complementary to the to the study of T2D as a binary disease state and as such we would hope to identify loci in our scans of FG and other traits which also convey risk for developing T2D. Given the modest sample sizes of our studies and the low population prevalence of T2D we do not have enough disease cases in our own sample data to draw conclusions about this. However it has been shown by larger meta-analyses, to which we contributed data, that many loci initially identified in analysis of FG are also strongly associated with T2D risk(Dupuis, Langenberg et al. 2010).

In the Dupuis analysis of 40,655 T2D cases and 87,022 controls 7 out of 17 loci which had been shown to have genome-wide significant associations with glycaemic phenotypes were also significantly associated with T2D as a binary disease trait. They also demonstrated that 11 out of 18 known T2D loci had at least nominally significant and directionally consistent association with FG(Dupuis, Langenberg et al. 2010).

Of particular note from the Dupuis analysis; the *MTNR1B* locus, which was strongly associated with FG in my own meta-analysis, had a significant risk for T2D with a relative risk of 1.09 for the FG raising allele. By contrast the most strongly associated FG signal in my analysis *G6PC2* gives a very weak and directionally inconsistent risk for T2D in the Dupuis analysis (RR=0.97, P=0.012). Dupuis noted that “a large T2D effect size does not always translate to an equivalently large FG

effect in non-diabetic persons"(Dupuis, Langenberg et al. 2010). This highlights the value of studying both case-control disease states and quantitative phenotypes.

5 Chapter 5: Genome-wide association of Pulse Wave traits in 2 isolated populations

5.1 Background to pulse wave analysis relating to cardiovascular disease.

As previously discussed CVD causes the largest proportion of deaths in the developed world and is estimated to be responsible for over 17 million deaths per year worldwide(WHO 2008).

Hypertension is a common condition estimated to affect approximately 26.4% of the adult population of the world(Kearney, Whelton et al. 2005). It is also known to be the strongest predictor of CVD development and outcome(He and Whelton 1997). Case control GWAS studies of hypertension and quantitative trait GWAS studies using systolic (SBP) and diastolic (DBP) blood pressure measures have been conducted in an attempt to identify genetic risk loci that influence control of BP and may play a role in the development of CVD (Cho, Go et al. 2009; Levy, Ehret et al. 2009; Newton-Cheh, Johnson et al. 2009; Org, Eyheramendy et al. 2009; Wang, O'Connell et al. 2009).

Results obtained in early GWAS studies had considerably less success in identifying loci for hypertension when compared to studies of other complex disorders of comparable size (WTCCC 2007) and similarly, studies using BP as a continuous outcome, which would be expected to have greater power than studies using a

categorical outcome measure, have required extremely large sample sizes to identify reliable association signals(Levy, Ehret et al. 2009). This may be due in part to the relatively poor reproducibility of BP measures when compared to, for example, blood biochemistry measurements(Crilly, Coch et al. 2007).

In addition to variation arising from measurement error BP is also susceptible to environmental influences that result in biological variability. It has been shown that factors such as recent smoking, needing to urinate or even verbal interaction during measurement result in substantial increases in BP measurements. Even in the most controlled environment some people react nervously to the clinical setting in which most BP measurements are taken creating a variable inter-individual "white coat effect" such that BP measurements from these settings do not reflect usual BP(Handler 2009). These additional sources of variance may obfuscate the relationships between long-term genetic influences on CVD and the snapshot measurement of BP.

It is also important to note that traditional measures of SBP and DBP give only an estimate of pressure in the peripheral arteries, SBP being a measure of the peak arterial pressure and DBP being a measure of the minimum peripheral pressure as the ventricles relax and the heart fills with blood. These peripheral pressure measurements may not accurately reflect the central arterial pressure particularly in the presence of decreased arterial elasticity (O'Rourke 1990; Papaioannou, Protogerou et al. 2009). When attempting to monitor or predict CVD risk, central

pressure is a stronger risk factor than peripheral BP and so an accurate estimate of central pressure should provide a better phenotype for identifying CVD risk genes.

One of the major causes of arterial stiffening is the formation of atheromatous plaques (atherosclerosis), which results in part from elevated levels of circulating LDL-cholesterol collecting in arterial lesions. As a major risk factor in CVD measures of blood lipid levels have been used extensively in GWAS analysis(Tohidi, Hatami et al. 2010).

More complex analysis of BP could theoretically prove more successful in identifying CVD risk genes by separating the observed variation into the individual components of cardiac function, valvular function and arterial function(Papaioannou, Protogerou et al. 2009).

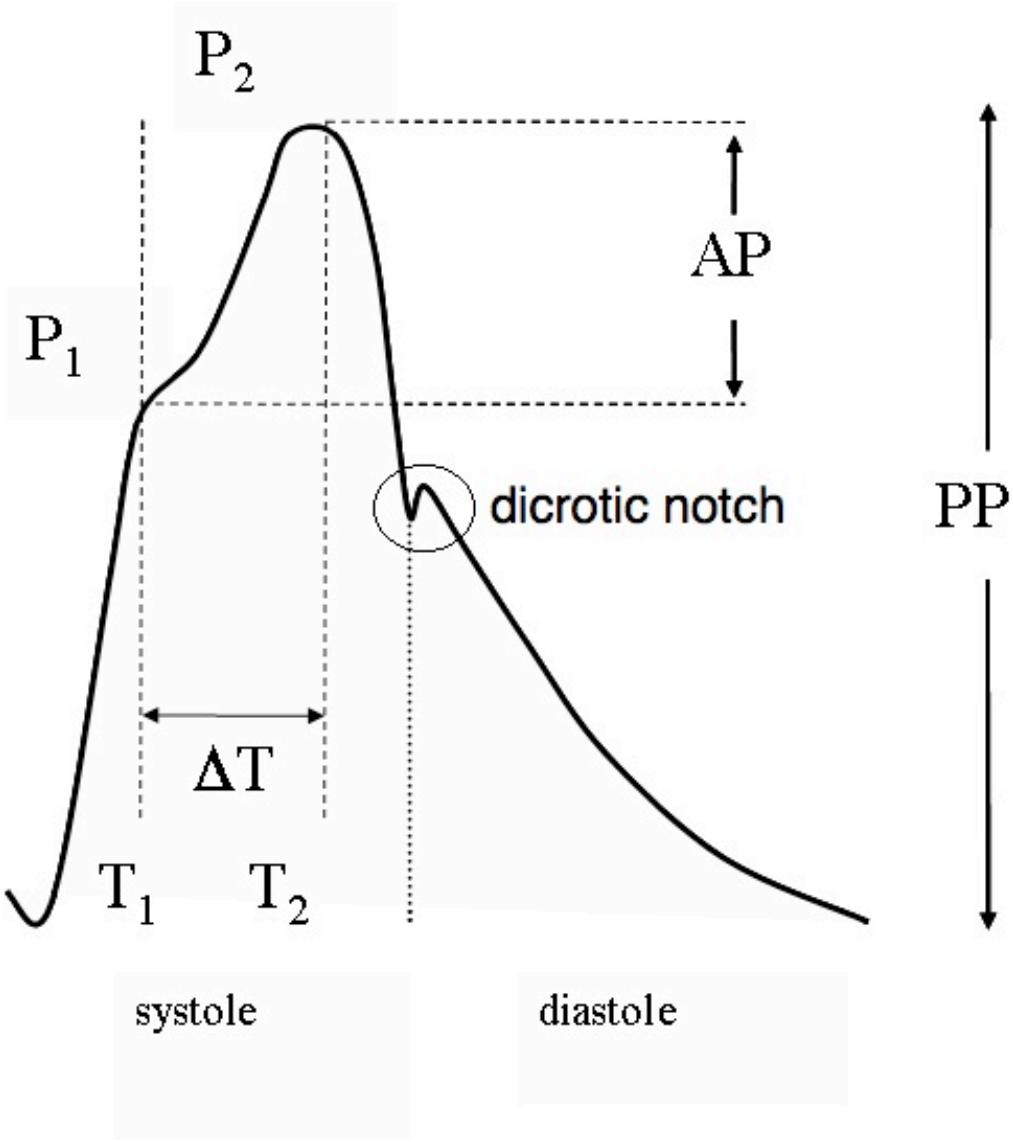
The two leading, non-invasive methods for measuring arterial stiffness are Pulse Wave Velocity (PWV) and Pulse Wave Analysis (PWA). PWV is calculated by detecting the arterial pulse wave at two points, for example at the carotid and femoral arteries. Taking the known distance between the two measurement points and dividing this by the measured time interval between the peak pressures at these two points gives the velocity of the pulse wave(Blacher, Asmar et al. 1999).

The PWA method uses a single measurement point at the radial artery(Papaioannou, Protogerou et al. 2009). A highly sensitive tonometry device is applied and measures

the changes in radial pulse pressure. The classical shape of the waveform, as shown in figure 27, is formed by the outgoing pulse wave and a weaker reflected wave as pressure in the peripheral arteries maximizes and the direction of flow reverses. The classical SBP and DBP measures represent the maximum and minimum of the waveform and in a healthy individual with extremely well functioning arteries these values should be proportional to the central arterial pressures. However, when there is significant arterial stiffness, the speed of the wave increases which in turn reduces the time until the reflected wave and thus increases the overlap between outgoing and reflected waves generating an augmentation of the peak pressure measurement(O'Rourke, Pauca et al. 2001).

The pressure waveform can be broken down into two time periods. The first is the systolic period, beginning when the left ventricle contracts and the aortic valve opens and ending when the aortic valve closes causing a small increase in arterial pressure known as the dicrotic notch(O'Rourke, Pauca et al. 2001). The diastolic period, during which the heart relaxes and fills with blood causing the arterial pressure to reduce, then lasts until the next pulse cycle.

Figure 27. Example pulse waveform describing key elements



There are several key elements of the pulse waveform from which our phenotypic data were derived. Firstly deriving the maximum slope at the start of the systolic pressure wave (MaxDp/dt) gives the maximum rate of pressure change during each

pulse and so can indicate problems with the functioning of the aortic valve(Starr, Ambrosi et al. 1973).

The ejection duration (EDA) is a measure of the systolic time period. At the end of the ejection period when the aortic valve closes and the dicrotic notch is observed the pressures within the left ventricle and arteries are approximately equal and so this ejection systolic pressure (ESP) gives an estimate of the ventricular pressure at the end of systole.

There are two peak pressure measurements P1 and P2 at time points T1 and T2 respectively. P1 is the peak of the outgoing pressure wave caused by ventricular contraction. P2 is the peak of the reflected wave. In a healthy individual P1 will be greater than P2 as the reflected wave will not occur until after the peak of the systolic pressure wave. However, when arterial stiffness results in an earlier stronger reflected wave the overlap will result in P2 becoming the maximum pressure measurement. A simplified measure of SBP cannot distinguish between these two cases. The pulse pressure PP is the difference between the maximum pressure measured during the pulse cycle, either P1 or P2, and the minimum diastolic measurement.

The interval between T1 and T2 (ΔT) estimates the time between the outgoing wave and the reflected wave and is proportional to the velocity of the pulse wave. This period will become shorter as the arteries become less flexible and BP rises.

The augmentation pressure (AP) is the difference between P1 and P2 representing a measure of the inflation of systolic pressure due to arterial stiffness. In healthy individuals the reflected wave should be delayed sufficiently that P2 is smaller than P1 and the AP is negative but in hypertensive individuals the rapidly and strongly reflected wave causes a substantial increase in pressure resulting in large positive values of AP. The clinical impact of this augmentation is relative to the overall pressure and so a commonly used alternative variable is the augmentation index (AIx), which is the AP expressed as a proportion of the PP. The AIx phenotype can be further refined by correcting for the differences resulting from heart rate (HR) and standardising to a HR of 75 beats per minute (AIx75).

Integration across the entire pulse waveform can give a measure of the overall mean arterial pressure (MeanP) indicating the general load on arterial tissues, while decomposition of the curve into the systolic and diastolic components to obtain a ratio gives the systolic ventricular index (SVI) also known as the Buckberg index.

Measuring theses phenotypes at the radial artery gives an estimation of peripheral arterial function, however the SphygmoCor PWA method allows for derivation of the central aortic waveform from which the same features can be calculated. The derivation of the central aortic waveform uses a general transfer function based on the experimentally defined relationship between radial and aortic pressures. The transfer function was originally calculated using simultaneous direct measurement of central aortic pressure and radial pressure in 20 patients undergoing cardiac

catheterisation surgery (Chen, Nevo et al. 1997). The general transfer function estimated from this study and implemented by the SphygmoCor system has been determined to be accurate by a subsequent replication study in which 62 cardiopulmonary bypass patients were assessed for directly measured aortic pulse waves and simultaneous derived central waveforms from radial tonometry (Pauca, O'Rourke et al. 2001).

The ability to derive central pressure waveforms is extremely important in relation to estimating CVD risk as the central pressure have been shown to be give far better prediction of cardiovascular disease outcomes than peripheral pressure measurements (Papaioannou, Protogerou et al. 2009).

To date only one genome-wide association (GWA) study has been performed using PWA phenotypes obtained using applanation tonometry. The Framingham heart study used 644 individuals and approximately 71,000 SNP markers tested against a battery of PWA ascertained from radial, brachial and femoral artery tonometry (Levy, Larson et al. 2007). The study identified several possible candidate regions, which showed weak association with one, or more of their tonometry phenotypes but none of the associations reached genome-wide significance or has yet been independently replicated.

Other studies of the genetics of arterial stiffness have focused on individual or small panels of candidate genes for which there is a biological link to arterial structure or

which have been previously implicated in cardiovascular disease studies. These studies have demonstrated association between a number of genes and one or more arterial stiffness phenotypes.

The genes encoding angiotensin (AGT), the angiotensin-converting enzyme (ACE), which converts angiotensin I to the more active angiotensin II, and the angiotensin receptor (AGTR1) have all shown association with arterial stiffness phenotypes, predominantly focusing on PWV, in candidate gene studies. This pathway has a known function in compensation for hypotension and polymorphisms in its component genes have been shown to interfere with the treatment of hypertension (Chung, Deng et al. 2009; Konoshita, Kato et al. 2009).

Other arterial stiffness studies have focused on structural elements with roles in cardiac tissue development. Some positive evidence for association has been found including elastin (Hanon, Luong et al. 2001), collagen (Brull, Murray et al. 2001)and fibrillin (Medley, Cole et al. 2002) genes.

While many of these candidate studies produced positive results, the evidence for association is relatively weak when compared to the levels required in a GWAS study and there are also a great many contradictory candidate gene publications showing no association with the same genes and phenotypes. For a full review see (Yasmin and O'Shaughnessy 2008).

We have used PWA to estimate the degree of arterial stiffness in healthy individuals from the ORCADES and KORČULA populations. The benefit of conducting genome-wide association is that no prior assumptions need to be made about the mechanisms underlying the disease and no prior knowledge of the function of the genes in question is needed. This allows for the identification of novel and unexpected candidates.

Our study has several benefits over the previous GWA study. Firstly we have performed meta-analysis of two distinct European populations. This provides internal replication of results, decreasing the chance of identifying false positives. Secondly we have genotyped over 300,000 single nucleotide polymorphisms (SNPs), providing considerably greater coverage of the genome. We have also conducted replication of a small number of the most significantly associated SNPs in a third independent cohort comprising 3,000 healthy British individuals.

Description of data collection

Pulse wave data were captured as part of a battery of tests in both study populations, for which participants had fasted for at least 2 hours. PWA was performed after 1 hour of supine rest. Radial arterial tonometry was performed using the SphygmoCor Px system (AtCor Medical, NSW, Australia) with a Miller tonometer (Miller instruments, TX, USA). Tonometry was performed continuously for 30 seconds in a resting state and an average waveform was generated. The analysis was performed a total of 4 times, two successful readings followed by 5 minutes rest followed by two

additional readings. Relevant measurements were derived from each waveform and the results from the 4 repetitions were averaged to give values for statistical analysis.

5.2 Meta-analysis results for ORCADES and KORČULA

5.2.1 Phenotypic Distributions

The two populations showed differences in the distributions of the majority of PWA phenotypes assessed (Table 15). Overall the differences correspond to a longer, slower pulse wave and lower pressure in the ORCADES population suggesting a lower degree of arterial stiffness in this population compared to the KORČULA population. This is not unexpected given that the ORCADES samples used in this analysis were on average 2.6 years younger than the KORČULA samples while also having lower average weight (-1.3 kg) and BMI (-0.2 kg/m^2). The ORCADES samples also had slower average heart rates (-3.4 bpm) and, as previously discussed the rates of diabetes and hyperglycaemia, both of which are known risk factors for CVD, in the study are also considerably higher than in the ORCADES population.

Differences of these kinds are not unexpected given the large distance, both geographically and perhaps culturally, between the two populations. Perhaps the most surprising thing is the direction of these differences. The traditional mediterannean diet has been cited numerous times as highly beneficial in preventing and controlling cardio-metabolic diseases (Esposito, Ciotola et al. 2007; Urquiaga, Echeverria et al. 2008; Babio, Bullo et al. 2009; Champagne 2009; Perez-Lopez, Chedraui et al. 2009; Proietti, del Balzo et al. 2009) while Scotland in general has

been suggested in the past to have an extremely poor diet in relation to CVD risk (Woodward and Tunstall-Pedoe 1995). The observed differences may suggest the lifestyle of Orcadians is healthier and unrepresentative of mainland Scottish populations or the difference in age and gender breakdown between the ORCADES and KORČULA studies may confound the expected health relationships.

Since the PWA traits we studied are derived from a single waveform there is a high degree of phenotypic correlation particularly between the measures of systolic pressure and augmentation. Age and sex adjusted correlation coefficients calculated using the ORCADES data for key traits are shown in table 14.

Table 14. PWA trait correlations in the ORCADES data

	P-MaxDp/dt	P-ΔT (ms)	P-SP (mmHg)	P-DP (mmHg)	C-AIx	C-PP (mmHg)	C-SVI
P-MaxDp/dt	1.00	0.24	0.73	0.13	-0.05	0.78	-0.21
P-ΔT (ms)	0.24	1.00	-0.06	-0.21	-0.23	0.03	0.19
P-SP (mmHg)	0.73	-0.06	1.00	0.67	0.18	0.80	-0.24
P-DP (mmHg)	0.13	-0.21	0.67	1.00	0.32	0.19	-0.15
C-AIx	-0.05	-0.23	0.18	0.32	1.00	0.33	0.22
C-PP (mmHg)	0.78	0.03	0.80	0.19	0.33	1.00	-0.07
C-SVI	-0.21	0.19	-0.24	-0.15	0.22	-0.07	1.00

Normal distributions were observed for the following phenotypes: AIx, the time of the reflected wave and the interval between systolic and reflected waves. All other phenotypes deviated from the normal distribution as assessed by the Shapiro-Wilks test. To allow for the accurate calculation of, and correction for, polygenic variance we used rank transformation to achieve normal distributions for all phenotypes.

5.2.2 Heritability estimates

Heritability estimates for many of the traits were different between the two study populations, with estimates from the ORCADES population generally being higher than those from the KORČULA population. However, standard errors in the KORČULA population estimates were high and many of the heritability estimates for this study are not statistically significant. There are several possible explanations for the differences shown. Firstly, accurate heritability estimates can only be obtained with a sufficient number of closely related individuals. For the purposes of PWA the sample size of the ORCADES study was approximately 19.7% larger than the KORČULA project. The level of kinship within the KORČULA study was lower, with average pairwise estimates of 0.0029 compared with the ORCADES estimate of 0.0046 indicating the presence of less family structure. Looking specifically at pairwise kinships above 0.35 that would suggest a full-sib or parent-offspring pairing we find 262 pairs in the KORČULA study and 547 in ORCADES, demonstrating a substantially higher level of close family structure. This suggests the ORCADES study would be expected to give considerably more reliable heritability estimates.

In addition to the size and genetic structure of the two studies we may also expect a greater deal of experimental variation in the KORČULA study because, while the ORCADES PWA measurements were all collected by, or under the observation of, a single individual who had many years of experience working with the equipment, the measures for KORČULA were collected by a number of field workers some of whom had only recently been trained in PWA.

Lastly, as with all the phenotypes measured in these studies, there may truly be greater levels of environmental variance in one population or the other resulting in a proportional reduction in genetic heritability.

Table 15. PWA trait descriptions and heritability estimates by population

	KORČULA					ORCADES				
	Mean	s.d.	h2	S.E.	p-value	Mean	s.d.	h2	S.E	p-value
N	487					583				
% Female	64					53.5				
Age (yrs)	56.2	13.9				53.6	15.7			
weight (kg)	79.2	14.4				77.9	15.3			
height (m)	1.68	0.09				1.67	0.09			
bmi (kg/m2)	28	4.15				27.8	4.86			
HR (bpm)	63.3	9.29	0.18	0.24	0.45	59.9	8.74	0.32	0.1	1.8x10-3
P-MaxDp/dt	862	320	0.33	0.13	0.01	809	228	0.22	0.1	0.02
P-T1 (ms)	110	14.6	0.54	0.16	7.6x10-4	112	17	0.04	0.05	0.42
P-T2 (ms)	222	21.4	0.12	0.1	0.24	234	19.9	0.24	0.08	4.4x10-3
P-ΔT (ms)	113	22.9	0.35	0.13	0.01	122	22.2	0.36	0.08	2.1x10-5

P-AIx	82.5	17.5	0.07	0.07	0.33	75.3	19.2	0.19	0.06	3.3x10-3
P-SP (mmHg)	138	21.8	0.33	0.15	0.03	130	17.9	0.21	0.08	0.01
P-DP (mmHg)	82.1	9.64	0.13	0.12	0.28	75.5	9.17	0.25	0.09	3.5x10-3
P-EDA (ms)	335	21.7	0.26	0.15	0.09	338	21.9	0.6	0.11	1.4x10-8
P-ESP (mmHg)	104	14	0.04	0.06	0.5	96	13.8	0.27	0.08	1.3x10-3
P-MeanP	102	13.4	0.09	0.1	0.37	93.8	12.3	0.25	0.09	3.5x10-3
C-T1 (ms)	106	10.3	0.24	0.14	0.1	111	11.7	0.28	0.09	1.5x10-3
C-T2 (ms)	234	18.4	0.34	0.23	0.13	235	20.5	0.21	0.08	0.01
C-ΔT (ms)	127	20.9	0.22	0.14	0.11	124	24.1	0.19	0.07	0.01
C-AIx	27.6	11.8	0.05	0.06	0.38	22.9	14.4	0.18	0.05	9.3x10-4
C-AIx75	22	11.3	0	0	0.5	15.6	14.4	0.22	0.08	3.5x10-3
C-AP (mmHg)	13.5	8.37	0.09	0.09	0.29	10.9	8.68	0.28	0.08	4.8x10-4
C-ESP (mmHg)	114	17.2	0.15	0.15	0.31	106	16.5	0.26	0.08	1.3x10-3
C-SP (mmHg)	128	21.5	0.25	0.15	0.11	118	19.1	0.24	0.08	3.5x10-3
C-PP (mmHg)	45.1	16.2	0.16	0.08	0.04	42	13.8	0.36	0.11	1.3x10-3
C-SVI	155	27.7	0.11	0.08	0.18	166	31.3	0.09	0.05	0.05

5.3 Association Analysis

Due to the relatively small number of samples from each population we focused only on the results of the meta-analysis of both studies giving a combined initial sample of up to 1060 individuals. This can also provide a form of internal replication as results showing a consistent direction and magnitude of effect in both study populations will have increased significance while false positives, arising from purely random correlation of genotype and phenotype in one study, are statistically unlikely to also be associated in the other study.

Meta-analysis of the two populations was performed for all subjects and separately for sex-specific sub groups. As with the glycaemic phenotypes, analysis was performed initially on rank-transformed phenotypes as the polygenic estimation method used to obtain residual phenotypic variance assumes a normally distributed phenotype. Subsequently the tests for significantly associated SNPs were repeated on untransformed variables to obtain effect estimates on a clinically meaningful scale.

Looking initially at the relatively simple and commonly used measures of HR and P-DP we can see from the QQ plots (figures 28 and 29) that the test statistics did not deviate greatly from the expected distribution under the null hypothesis. In contrast the P-DP analysis shows an inflation of the test statistics at the $1x10^{-5}$ level. From the genome-wide association plot (figure 30) we can see that the four most highly significant results for P-DP fall within the same region of chromosome 9. The strongest associated marker, rs574090 ($p=2.3 \times 10^{-7}$), is situated in an intron of the gene *ADAMTSL1*.

Figure 28. QQ plot of heart rate GWAS meta-analysis results in ORCADES and KORCULA studies

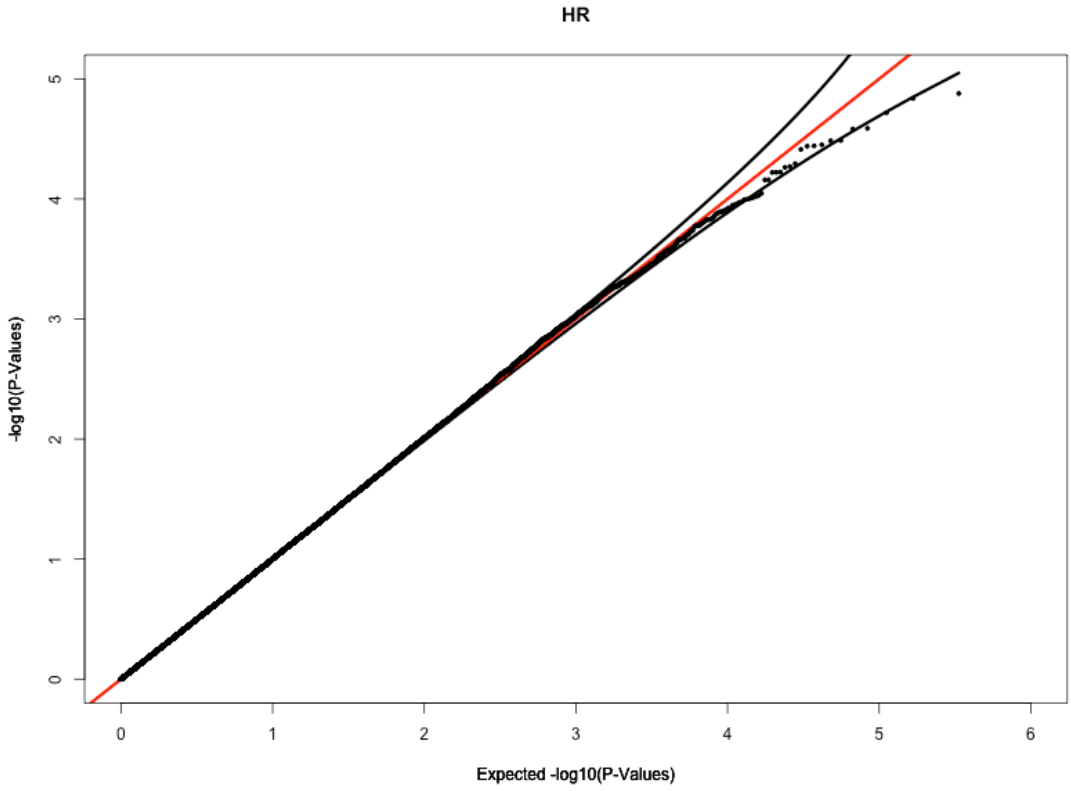


Figure 29. QQ plot of peripheral diastolic pressure GWAS meta-analysis results in ORCADES and KORCULA studies

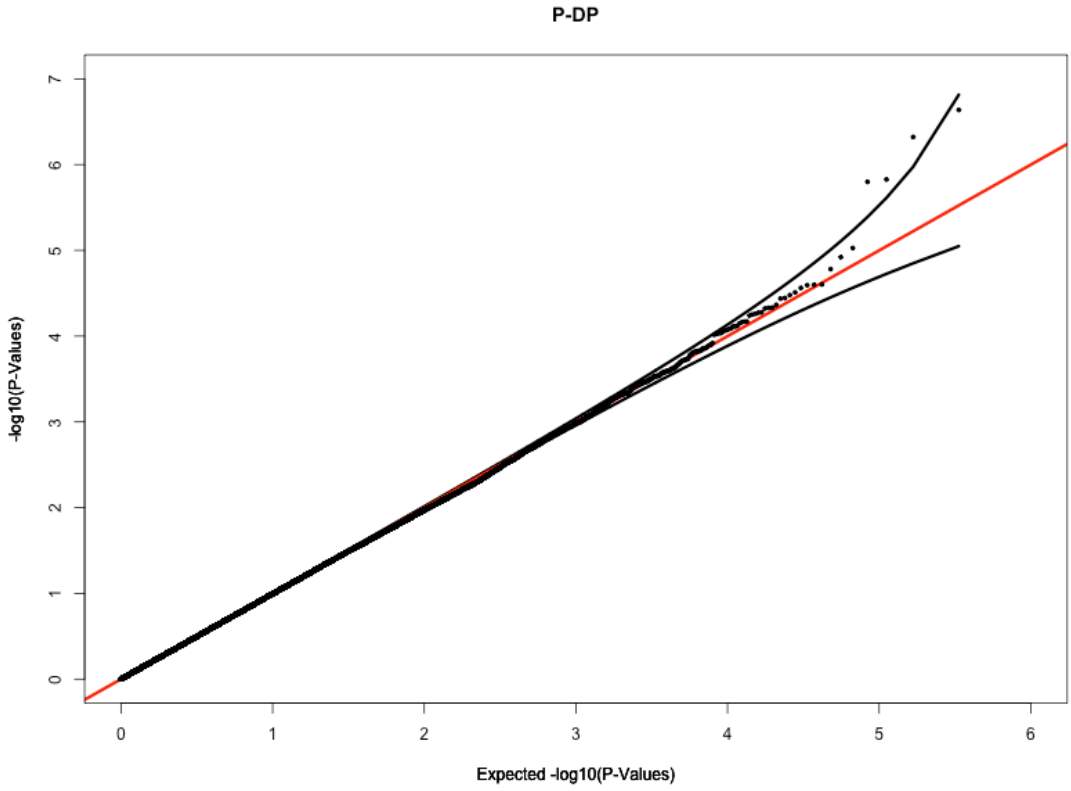
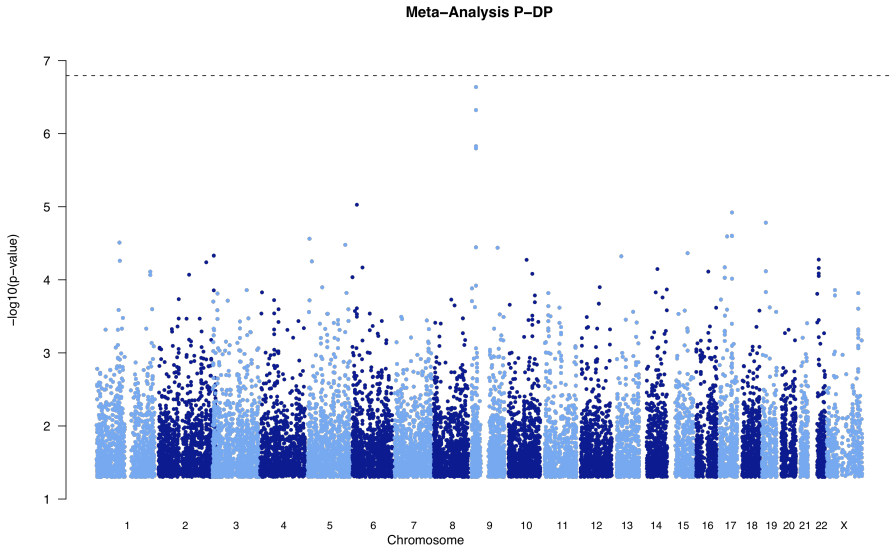


Figure 30. GWAS meta-analysis results for peripheral diastolic pressure



The QQ plots for C-AIx and the heart rate adjusted C-AIx75 both show a certain degree of inflation at the higher test statistics with the effect being more pronounced in the C-AIx75 results. This may indicate that, as we hoped, removing the "noisy" variance in augmentation caused by variable heart rate allows for better detection of genetic sources of variance.

The two association scans show a potential region of interest on chromosome 16. The strongest association is with SNP rs4843294 while several other markers in the region show weaker levels of association. This result is stronger in the heart rate corrected C-AIx75 analysis with the p-value moving from 7.0×10^{-7} in C-AIx to 2.2×10^{-7} in C-AIx75. The peak is located within an intron of the *BANP* gene.

Figure 31. QQ plot of central augmentation index GWAS meta-analysis results

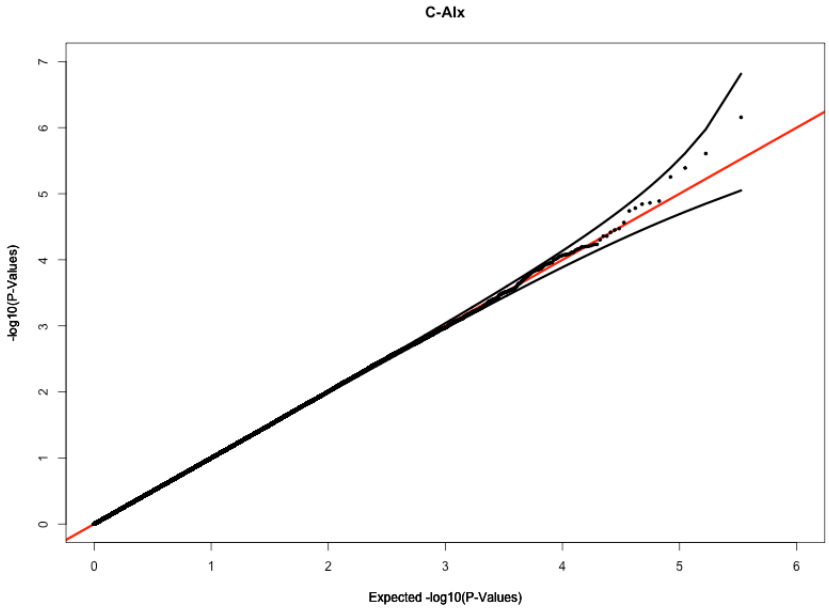


Figure 32. QQ plot of heart rate corrected central augmentation index GWAS meta-analysis results

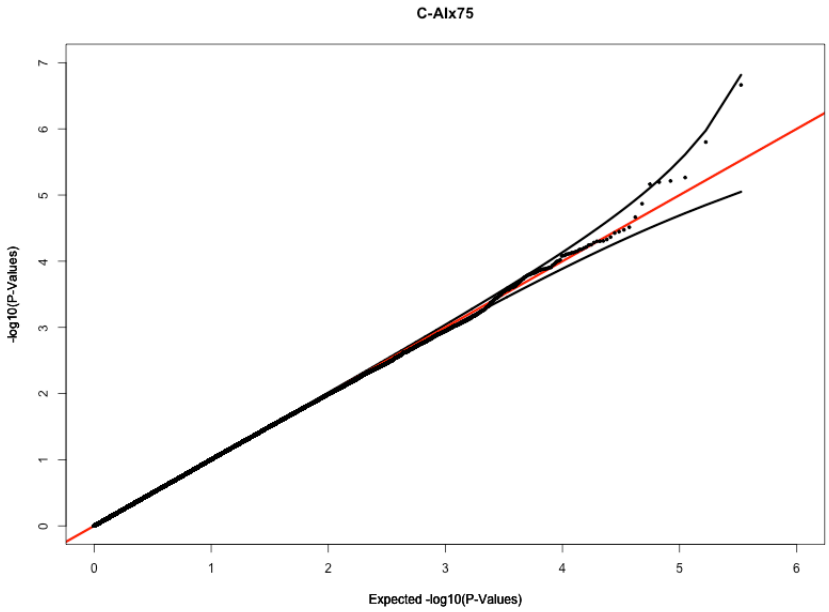


Figure 33. GWAS meta-analysis results for central augmentation index

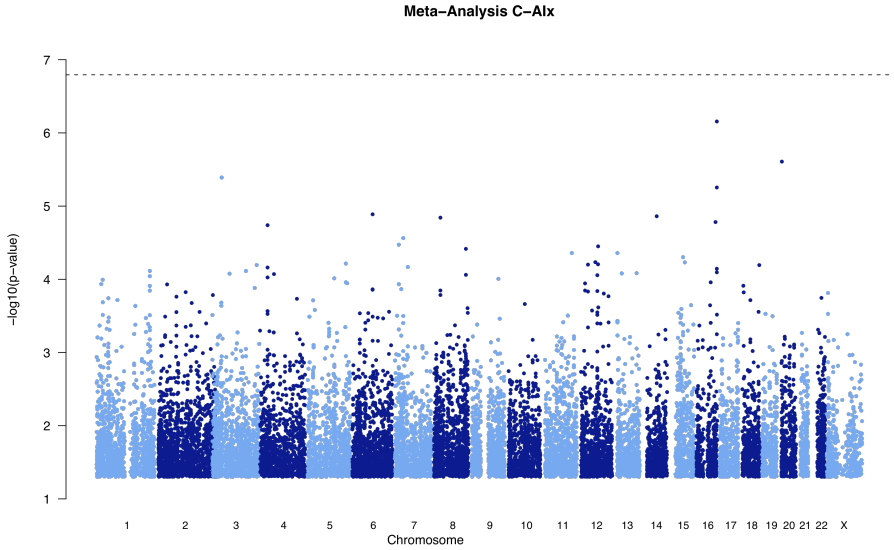
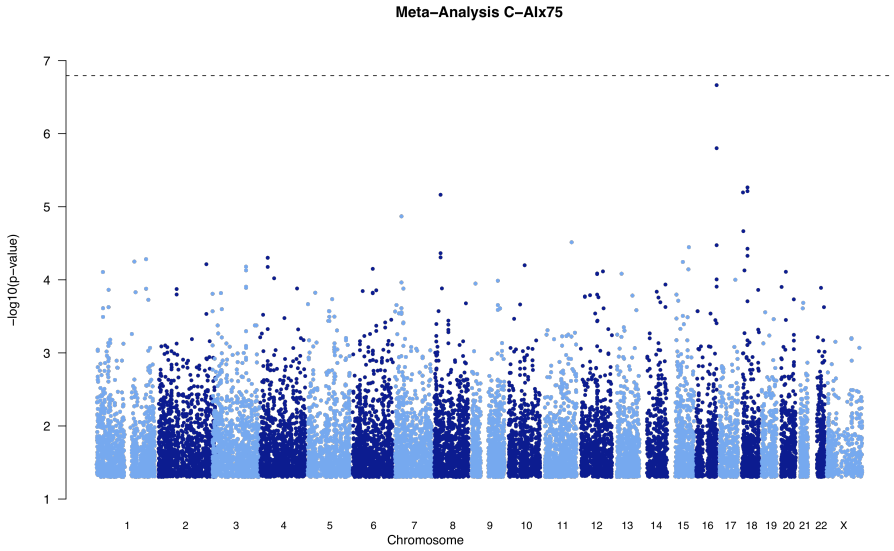


Figure 34. GWAS meta-analysis results for heart rate corrected central augmentation index



The wave interval measure C-ΔT gave a near genome-wide significant result on chromosome 21 (rs9981633, $P=2.4x10^{-7}$).

Figure 35. GWAS meta-analysis results for time interval between outgoing and reflected pressure waves

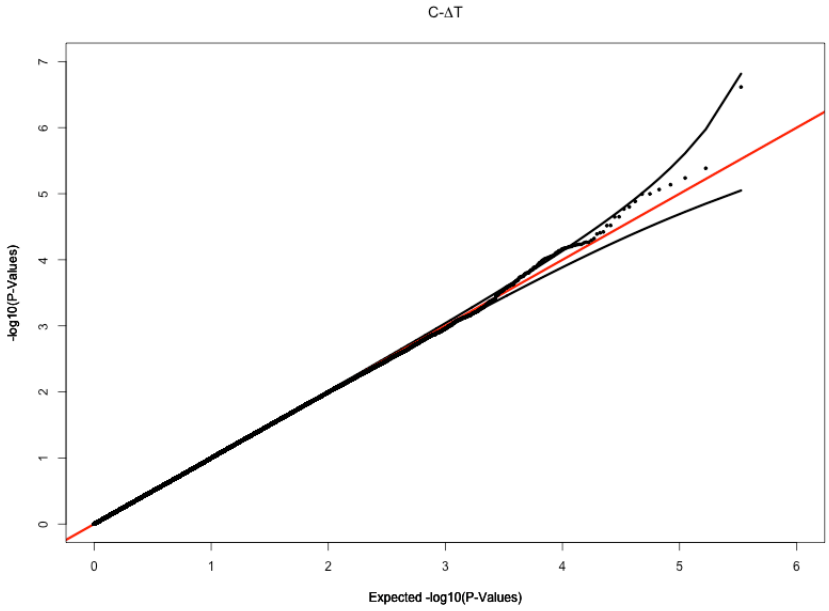
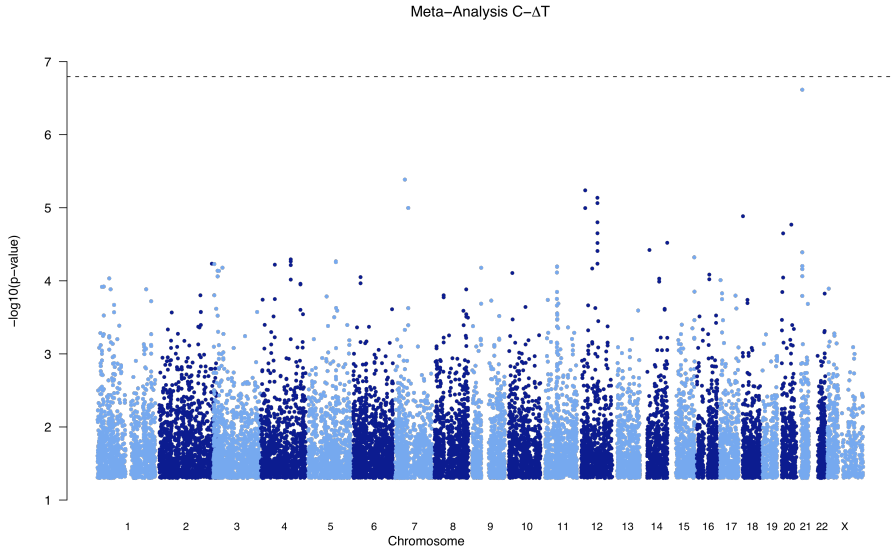


Figure 36. QQ plot of heart rate corrected central augmentation index GWAS meta-analysis results



Several of the most promising markers, chosen based on strength of association and genomic context, were genotyped in an independent sample of ~3000 healthy individuals from the Anglo-Cardiff Collaborative Trial (ACCT) who had been previously measured for some of the PWA phenotypes. The results (table 16) show replication at a nominal P-value of 0.01 for the rs574090 association with P-DP but unfortunately fail to provide replication for any of the derived central pulse wave phenotypes.

Table 16. Discovery meta-analysis and ACCT replication results for selected PWA SNPs

SNP	Trait	N	Freq	Effect	S.E.	P Discovery	Effect ACCT	S.E. ACCT	P ACCT
rs4843294	C-AIx	1069	0.74	-2.29	0.44	$2.2x10^{-07}$	-0.5	0.47	0.28
rs574090	P-DP	1075	0.88	-3.18	0.62	$2.3x10^{-07}$	1.87	0.72	0.01
rs4793570	C-ESP	1075	0.27	-3.05	0.67	$5.5x10^{-06}$	-0.63	0.86	0.46
rs9304718	C-SP	1075	0.36	2.95	0.74	$6.4x10^{-05}$	0.67	0.88	0.45

With only four markers and a prior expectation of association there is no way to identify or correct for inflation of the test statistics in the ACCT study. However since these samples were randomly selected from across England and Wales and consist of unrelated individuals we would not expect significant inflation to be present and, in the absence of any strongly replicating marker we can safely assume that inflation is not an issue with these results.

5.4 Conclusion

The initial results from the two-population GWAS meta-analysis provided several potential novel associations. However, as the results did not reach genome-wide significance in the initial analysis and given the lack of replication for these results in the considerably larger ACCT study we may conclude that the results are most likely false positive associations.

The positive inflation in test statistics demonstrated by the QQ-plots for several of the traits may indicate there are genuine genetic factors that could be identified from these populations but a large scale replication effort including several hundred of the top markers may be necessary to distinguish between the true and false positives.

Increasing the initial GWAS sample size would also have major benefits in increasing the power to detect real associations at a given significance threshold. In this respect planned expansion of the ORCADES study and addition of other cohorts with PWA and genome-wide SNP data would both greatly enhance the discovery cohort size.

6 Chapter 6: Discussion

This project had two intrinsically linked aims. One was to attempt to identify novel quantitative trait loci that were involved in the control of metabolic and cardiovascular phenotypes and which may therefore be related to the common complex diseases of T2D and CVD. The second was to explore the specialized methodologies required to identify such QTLs using isolated population studies.

6.1 Linkage

In respect of the second aim I initially attempted linkage analysis of FG levels, as this was by far the mostly widely available phenotype in our study populations. The results of this analysis were not particularly strong, with no peak reaching the significance level for our meta-analysis and only one peak, on chromosome 17 in ERF, reaching the individual population significance threshold. Two of the sub-significant peaks I detected did overlap with previously reported linkage results for glucose or T2D studies suggesting they may be more likely to be genuine linkage signals.

These possible linkage signals on their own do not yield a great deal of information about possible FG related genes. The vast width of the peaks I detected contain hundred of possible genes, any of which could theoretically be the source of the linkage signal.

To attempt to narrow down the field of targets and to provide some degree of confirmation of the linkage signals, single point association analysis was done using the SNP data available for each cohort. These association results did not provide any additional support for the majority of the linkage peaks. The peak on chromosome 17 did show some degree of association. In both the linkage and association analysis this result was only seen in the ERF population. This could indicate the presence of a population specific polymorphism with a strong effect on FG levels but with no visible effect in the other EUROSPAN populations this is a difficult hypothesis to pursue.

The linkage analysis had several limitations, each of which probably contributed to the lack of substantial results. Firstly the studies used were not designed specifically to recruit families and so the level and nature of pedigree structure varies from one study to the next. This makes it difficult to estimate the power for linkage analysis within these populations. However, given the inconsistent results obtained between the four study populations and the weak overall results obtained from meta-analysis it may be safe to conclude that they were not sufficiently powered for analysis of a phenotype as highly polygenic as FG.

The second problematic element of the linkage analysis was the fact that each of the studies had been genotyped using a different set of genetic markers and in fact the ERF study had used an entirely different class of markers to the other EUROSPAN

studies. Additionally, the NSPHS microsatellite marker positions were reported on a different genetic map to the other datasets.

While I and the other EUROSPAN analysts attempted to overcome these differences by interpolating the NSPHS map positions and calculating MIBD estimates at arbitrary intervals along chromosomes rather than at individual marker positions, the resulting meta-analysis alignments may not be as accurate as would have been possible if a consensus marker panel had been typed in all populations.

Besides any drawbacks with the study design, the linkage method is inherently limited to detecting rare variants with large individual effects. Given what we now know about the genetics of FG, with 16 replicated variants each of which exerts only a small influence on FG levels (Dupuis, Langenberg et al. 2010), the linkage method may not be an appropriate choice for studies of this type of highly polygenic phenotype.

6.2 Association methods

While the family structure of our isolated populations had potential benefits for linkage mapping, it was simultaneously troublesome for the purposes of basic association analysis. It became clear early on in the project that the widely used GC correction method would not be appropriate or sufficient to correct for the high levels of inflation seen in our GWAS statistics.

The use of pedigree-based kinship correction to obtain residual phenotypes under the GRAMMAR model overcame the family-based inflation in a reasonably accurate fashion. It did however introduce its own problems in that it required complete pedigree information, which in human populations is all but impossible to obtain. My implementation of the GRAMMAR method also required the polygenic correction and association analysis steps to be performed in separate programs using separate data formats, thus increasing the chances of user-generated errors.

The GRAMMAR method may have been acceptable for use in the EUROSPAN populations, all of which had extensive pedigree information. However with the addition of the KORČULA study, which was prediced to have a substantial degree of family stucture but for which no deep pedigree information was avaiable, an alternative method would have been required.

The third possible correction method I investigated was the use of genomic based kinship estimates as implemented in the GenABEL analysis package. The use of genomic rather than pedigree-based kinship overcame many of the problems and pitfalls of the GRAMMAR protocol. Most importantly for the KORČULA analysis was the fact that no pedigree information was required and the method can be used in any population for which significant population structure is predicted.

More widely the method is beneficial even where some pedigree information has been collected since it does not really on the accuracy or completeness of this data. I

would also suggest that, even in the hypothetical scenario of having a complete and accurate pedigree, the genomic kinship estimate is more useful for the correction of polygenic trait variance because it accounts for the stochastic variance brought on by semi-random meotic events.

Lastly I would argue the benefits of conducting both the polygenic phenotype correction and association analysis within a single program. This was not only faster and more convenient for the analyst but also reduced the chances of simple but difficult to identify mistakes being made as the data were reformatted and transferred.

6.3 Glycaemic association results

In the respect of identifying novel QTLs, the most significant success from the project was the identification of the association between the *TCF7L2* locus and HbA_{1C} variation in healthy individuals. While this locus has been previously and repeatedly shown to be associated with T2D it had not been previously demonstrated to influence HbA_{1C} or, until very recently(Dupuis, Langenberg et al. 2010), FG levels in non-diabetic individuals. This result may therefore contribute additional information about the mechanism by which *TCF7L2* polymorphisms act to increase T2D risk.

The *TCF7L2* locus has not been identified as a significant factor in other GWAS studies of HbA_{1C} and a recently published large-scale meta-analysis of the trait also

failed to show a strong effect for this locus (Soranzo, Sanna et al. 2010). This repeated failure to replicate would usually strongly suggest that the result observed in my analysis is a false positive. However there is a strong biological relationship between HbA_{1C} and T2D and given that *TCF7L2* is an extremely well replicated T2D locus the question of whether my result is genuine requires closer examination.

Arguing from the position that the result is false, the first argument one could make is that the association, like many other non-replicated results, is a statistical artefact arising by pure chance. Using a Bonferroni correction of the P-value of 1.48x10-7 for 256,587 SNPs included in the HbA_{1C} meta-analysis the empirical P-value is just 0.038. If I also took into account the various other GWAS scans I perform on various phenotypes then some would argue we should expect to find one or more results at this level.

In response I would argue that the probability that the marker showing the most significant result in an analysis of 256,587 SNPs would also happen to be widely reported in analysis of a strongly related disease is extremely small.

The relationship between the HbA_{1C} phenotype and T2D status leads to a second possible argument against the validity of the result. The HbA_{1C} measurements of T2D patients are generally considerably higher than those of healthy individuals. We would therefore expect to find, in a population containing significant numbers of T2D sufferers, some degree of association between HbA_{1C} and any sufficiently

strong T2D locus. Despite my efforts to exclude people with diabetes from analysis this phenomenon could still be an issue.

Due to the variable types of data available for the various cohorts used in my meta-analyses, the two exclusion criteria used were self-reported anti-diabetic medications and hyperglycaemia based on FG levels. This means that hypothetically; a diet controlled T2D patient who had successfully reduced their FG levels below 7mmol/L would be retained in our analysis. I do not believe this to be the foundation of the *TCF7L2* association, as I would only expect a very small proportion of individuals to fall into this category.

There are also several explanations as to why the association may not be observed in other published studies despite being a genuine result. Firstly, one of the original rationales for using isolated populations was the relative reduction in genetic heterogeneity and environmental variance compared with urban populations. Both of these factors generate phenotypic "noise" and may potentially mask the relatively small effect of an individual QTL.

Secondly we must consider the power of the studies. It is not unusual or unexpected for relatively small studies, each of which samples a few thousand individuals, to give conflicting results. The seemingly large phenotypic effect of the *TCF7L2* variant in my analysis may suggest that relatively small sample sizes would be required to replicate the finding. However, effect estimates from studies that first

identify a new locus are generally higher than estimates from subsequent replication studies, a phenomenon known as the winners curse (Kraft 2008). This means the true effect of this variant on HbA_{1C} levels is likely to be considerably smaller than it appears in my analysis, and so this locus could easily be missed in other relatively small studies.

The largest published GWAS study of HbA_{1C} by the MAGIC consortium totaled almost 36 thousand individuals and was estimated to have 80% power to detect variants responsible for as little as 0.12% of total phenotypic variance (Soranzo, Sanna et al. 2010). While this type of large collaboration may give increased power it also creates certain issues.

When combining the analysis of many studies conducted by different research groups there are a variety of sources of heterogeneity and potential error. As observed in the relatively small EUROSPAN meta-analysis there may be different sources of environmental variance on any given phenotype in different populations. Secondly the experimental protocols used to measure phenotypes and collect relevant covariates may be different, and have differing levels of accuracy, between studies. There is also an issue when conducting a large meta-analysis of unified analysis protocols. Depending on the structure of the collaboration the analysis of each individual study may be conducted by a separate individual and, while all analysts may be following the same protocol, this means a mistake by any one of dozens of individuals could adversely affect results. These factors can all lead to increased

levels of environmental or error variance and may mask the effects of underlying genetic loci.

There is also a potential issue of genetic heterogeneity. Some genetic variants may be far less frequent or even absent in certain populations due to historical population bottlenecks or genetic drift. The addition of these study populations to a meta-analysis would therefore dilute the evidence for that particular variant.

The LD structure can also vary from one population to another and this can have major implications for the association analysis method. Since we do not necessarily expect the SNP variants we are testing in GWAS to be the actual causal variant we rely on the degree and direction of genotypic correlation between our SNP and the unobserved locus being the same in our separate study populations. If this LD relationship had been broken down in a particular study then the association with our SNP would no longer be present and, in a more extreme scenario, if the LD relationship had been reversed in a particular population, which may occur during a population bottleneck, that evidence from that study population would be contradictory and reduce the overall meta-analysis statistic by a substantial amount.

In addition to the identification of a new HbA_{1C} association my meta-analysis of FG levels also showed replication to varying degrees of 9 out of 16 published loci, which were tagged by SNPs on our marker panel. Though it must be noted that analysis of the EUROSPAN populations was also included in one of the major FG meta-analysis

(Dupuis, Langenberg et al. 2010). Additionally 1 of 2 well-replicated FI loci, the insulin like growth factor 1 (*IGF1*), showed nominal association in our data (p=0.006).

At the same time it is interesting that a further 7 FG and the remaining 1 FI loci showed no evidence for association in our study populations. As already mentioned the power for our study to identify small effect loci at genome-wide significance levels is quite low and the limited number of studies for which FI was measured decreases the power further in that analysis. However, even for a locus whose effect accounted for 0.5% of total phenotypic variance, we would expect to have more than 80% power to detect the variant at a p-value of 0.001 (Purcell, Cherny et al. 2003), using our 4,000 sample FG meta-analysis. This suggests that some of the loci identified by large-scale multinational meta-analyses may not be having an effect in our population isolates. There are many plausible explanations for how this could occur.

While we expect the overall LD in our isolated populations to be stronger than that found in larger populations, it is also plausible that different patterns of LD could have developed in some of our study populations. This may mean that markers that are in LD with a causative variant in European populations at large do not show the same correlation in our studies. The possibility of discrepancies between LD structure in our studies and that found in the CEU HapMap samples is difficult to

assess, as we do not have substantial numbers of founder individuals with which to calculate LD.

Another possible explanation for our inability to replicate some of the major FG loci is that population bottlenecks and low effective population sizes increase the chance of random genetic drift changing allele frequencies or even taking them to fixation in the population. While this would be easily observable at directly genotyped variants, it could result in unobserved causative variants being present at very different frequencies, or not present at all, in our study populations.

A further consideration is the possibility of currently unknown genetic or environmental modifiers that are necessary for a given locus to influence a trait. In the case of environmental factors this could be a result of the dietary intake of some of our study populations being generally healthier. Many loci have been identified recently with significant effects on obesity (Fawcett and Barroso 2010) but arguably they only show such strong influences in populations with access to large quantities of sugar and fat rich food. In the example of genetic modifiers the influence of a locus on a given phenotype may be dependent on the presence or absence of a second variant. The disentanglement of this type of interaction relationship is extremely difficult as, theoretically, any two loci in the genome could be interacting. Testing all possible combinations of SNPs instantly turns the millions of tests that are common in GWAS studies into trillions of tests, and the type I error rate increases accordingly.

The effect size estimates for loci that have been consistently associated with FG for example range between 0.012 (*IGF1*) and 0.075 (*G6PC2*) mmol/l per allele, and even collectively the 16 strongest known FG loci are only estimated to account for approximately 3-4% of total phenotypic variance (Dupuis, Langenberg et al. 2010).

So we can see that an individual study such as the ORCADES project with approximately 700 genotyped and phenotyped samples will have very limited power to identify even the strongest of these loci.

Expanding to a relatively small collaboration like EUROSPAN, which with the addition of the KORČULA project gave a total sample size of approximately 4000 samples for FG analysis, greatly increases our power to detect loci with effects similar to the known FG loci but simultaneously introduces problems of inter study genetic and phenotypic heterogeneity.

To further address the issue of power and extract as much value as possible from the EUROSPAN data sets we contributed analysis of FG, HbA_{1C}, Insulin and the HOMA-B HOMA-IR phenotypes to the meta-analysis of glucose and insulin related traits consortium (MAGIC) (Dupuis, Langenberg et al. 2010). Having collected together 21 GWAS studies and a large number of suitable replication samples the MAGIC consortium were about to reach sample sizes of between 98,372 and

122,743 for the various glycaemic phenotypes. It is at these levels that power exists to reliably identify loci with very small effect sizes.

Despite only having small numbers of type 2 diabetes cases roughly proportional to the population prevalence we were also able to contribute case control analysis results to a large scale meta-analysis (Voight, Scott et al. 2010). The small number of cases within our own studies made independent case control analysis largely redundant and, as previously discussed, methods to correct for the effects of diabetes medication on the glycaemic phenotypes were not applicable when such small numbers were available to calculate the medication effect estimates. This means that joining the large-scale meta-analysis made use of genotyped individuals who would not otherwise have been contributed to diabetes and related traits research.

6.4 Pulse wave analysis

Use of the PWA phenotypes had potential benefits in that this measurement is so difficult and time-consuming to obtain that very few genetic studies have used it to date. Unfortunately this also meant that only two of the study populations had data for PWA and the numbers of measured individuals within these studies were lower than for the more common biochemical and anthropometric phenotypes.

The estimation of PWA heritability in the ORCADES and KORČULA studies showed very inconsistent results. The KORČULA estimates were generally not statistically significant due to high standard errors. This could either be the result of

increased measurement error in therefore proportionally decreased genetic variance, or it could relate to the lower levels of kinship present in the KORČULA population and a reduced ability to estimate heritability.

In terms of association results the PWA did not reveal any convincing candidate loci. The strongest result from across the analysis was the association between marker rs574090 and peripheral diastolic pressure. This SNP also showed a nominal significance in the ACCT replication samples, the effect here however was in the opposite direction to the ORCADES and KORČULA populations.

The issues of power as discussed in relation to the glycaemic phenotypes are even more relevent to the lack of PWA association hits. With a sample size of little over 1,000 individuals across the two meta-analysis populations we would only have 80% power to detect very strong variants accounting for more than 4% of total phenotypic variance. Given the difficulties of obtaining accurate and consistent blood pressure measurements as previously discussed it is highly unlikely that a common genetic variant would account for that proportion of variance.

6.5 Extensions and alternatives to genome-wide association analysis

The simplest strategy for increasing power and improving results from GWAS studies is to increase the number of samples. This can be achieved, as in the

EUROSPAN, MAGIC and other collaborations, by combining analysis from many individual studies, though this type of meta-analysis comes with many hurdles and limitations as previously discussed. It would clearly be preferable to collect very large samples within a single study framework, which would allow for true standardization of data collection and analysis methods. The potential for this is obviously limited by the size of available research grants, but as the cost of genotyping technology has and continues to lower the feasibility of collecting and genotyping tens of thousands of samples within a single study increases.

In the absence of a very large samples size, one alternative economical method for improving study power is to implement a two-stage study design. The first stage is to type a genome-wide SNP panel on a subset of samples and conduct a standard GWAS analysis. The second stage is then essentially a large-scale replication using the remaining samples. Rather than choosing only a small set of SNPs with genome-wide significant P-values one can, with sufficient funds, take forward several thousand markers to the second stage chosen either using an arbitrary significance threshold or by taking a certain proportion of the most significant SNPs.

This two-step study design differs from more common replication efforts in that the SNPs used for replication, or at least the majority of them, are not expected to successfully replicate. Assuming the first stage study were to use the same genotype chip used in the EUROSPAN analysis with roughly 300,000 SNPs and an arbitrary P-value cut off of 0.001 one would expect to find approximately 300 SNPs that have

no genuine role in the phenotype being studied. Analyses that find a large excess of results at this level but which do not identify genome-wide significant results may indicated that a large number of weak effects are present. One potential replication strategy would therefore be to identify a significance threshold in your first stage study at which an excess of associations are found and then attempt replication of all markers above this threshold.

The ability to identify true QTL signals is also dependent on having sufficient coverage of the genome and the Illumina 317k chip used in the EUROSPAN population has been estimated to have as little as 59% coverage in European populations (Wollstein, Herrmann et al. 2007). While the density of SNP arrays has increased during the course of this project and chips with over 1 million SNPs are now available, complete coverage based on LD is almost impossible to achieve due to mutation hotspots (Tian, Wang et al. 2008) and variable recombination rates across the genome (Gay, Myers et al. 2007). The only sure-fire way of identifying all genetic variants an individual is carrying is through sequencing, a process that is still relatively time consuming and expensive when compared to micro-array based genotyping.

While whole genome sequencing may not yet be feasible for most research groups, collaborative projects are underway to sequence and make available the genomes of thousands of individuals (www.1000genomes.org). The availability of these

sequence data will enable groups with sufficient genotype data to impute the vast majority of common genetic variants present in their population.

Application of this whole sequence imputation to isolated populations such as ORCADES would have the same caveats as previous imputation methods in that differences in LD structure and genetic drift would result in lower accuracy. However, isolated populations also have a potential advantage in this respect. Thanks to the extensive family structure present in ORCADES and similar studies a relatively small number of sequenced samples could be used to accurately impute and phase the whole genome sequence of hundreds of genotyped samples. This pedigree based sequence imputation means that not only should it be possible to identify the actual causal variant for a given phenotype or disease, but also to trace parent of origin for each allele.

There is another likely outcome of the decreasing cost and increasing availability of sequencing technologies in the reversal of some traditional selection strategies. One of the proposed benefits of the isolated population study design was the ability to collect a large number of phenotypes all of which could be tested using a single set of genotypes. This strategy is indeed highly efficient when the cost of genotyping is very high and the cost of phenotyping is reasonably low. With a similar purpose, some studies have previously used selective genotyping strategies whereby a large number of individuals are phenotyped for a quantitative trait, but only the tail ends of

the distribution are sent on for genotyping in an attempt to maximize the genetic differences between them.

As high density genotyping and sequencing become ever cheaper, and geneticists attempt to tackle ever more complex and intensive phenotypes such as PWA, the economy of these types of strategies decreases. A situation is feasible whereby instead of phenotyping a large number of individuals and choosing a subset for genotyping, one could genotype a very large number of unselected individuals and choose a subset from this pool to send forward for intensive phenotyping. This strategy would obviously require a return to candidate gene selection in order to choose the phenotyping pool, but in light of several years of successful GWAS studies the knowledge on which to choose one's candidates is greatly increased.

The strategy would particularly benefit the study of rare variants. Given a very large sample of say 100,000 genotyped individuals, a variant with a minor allele frequency of 1% would be predicted to occur in approximately 1,000 of those samples. Assuming this low frequency variant was of interest one could preferentially approach those 1,000 individuals plus 1,000 individuals chosen randomly from the remaining pool for phenotyping, one could then test for the hypothesized effect of the rare variant with substantial power.

In addition to new methods to identify risk loci it is important to translate findings into clinically meaningful resources. With respect to the findings presented in this

thesis; the *TCF7L2* locus was first shown as a risk locus for T2D(Grant, Thorleifsson et al. 2006), it has since been shown to influence insulin secretion(da Silva Xavier, Loder et al. 2009) and from my own work to play a role in control of HbA1c levels in the non-diabetic population(Franklin, Aulchenko et al. 2010). The risk allele at this locus has been shown to not only increase lifetime risk of developing T2D but also is also associated with the age-of-onset for the disease(Lehman, Hunt et al. 2007; Silbernagel, Renner et al. 2011). Perhaps the most clinically tractable finding to date with regard to the *TCF7L2* locus is the reported and replicated finding that variation at this locus affects response to sulfonylureas a common form of treatment for T2D patients(Pearson, Donnelly et al. 2007; Holstein, Hahn et al. 2011). Findings of this type mean that the wealth of genetic information that has been gathered through genome-wide investigations of the last few years will increasingly find its way into the clinical world.

7 Bibliography

(1986). "The Diabetes Control and Complications Trial (DCCT). Design and methodologic considerations for the feasibility phase. The DCCT Research Group." Diabetes **35**(5): 530-545.

Abecasis, G. R., S. S. Cherny, et al. (2002). "Merlin--rapid analysis of dense genetic maps using sparse gene flow trees." Nat Genet **30**(1): 97-101.

Allison, D. B. (1997). "Transmission-disequilibrium tests for quantitative traits." Am J Hum Genet **60**(3): 676-690.

Almasy, L. and J. Blangero (1998). "Multipoint quantitative-trait linkage analysis in general pedigrees." Am J Hum Genet **62**(5): 1198-1211.

An, P., B. I. Freedman, et al. (2005). "Genome-wide linkage scans for fasting glucose, insulin, and insulin resistance in the National Heart, Lung, and Blood Institute Family Blood Pressure Program: evidence of linkages to chromosome 7q36 and 19q13 from meta-analysis." Diabetes **54**(3): 909-914.

Anderson, C. A., F. H. Pettersson, et al. (2010). "Data quality control in genetic case-control association studies." Nat Protoc **5**(9): 1564-1573.

Aulchenko, Y. S., D. J. de Koning, et al. (2007). "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis." Genetics **177**(1): 577-585.

Aulchenko, Y. S., P. Heutink, et al. (2004). "Linkage disequilibrium in young genetically isolated Dutch population." Eur J Hum Genet **12**(7): 527-534.

Aulchenko, Y. S., S. Ripke, et al. (2007). "GenABEL: an R library for genome-wide association analysis." Bioinformatics **23**(10): 1294-1296.

Babio, N., M. Bullo, et al. (2009). "Mediterranean diet and metabolic syndrome: the evidence." Public Health Nutr **12**(9A): 1607-1617.

Beer, N. L., N. D. Tribble, et al. (2009). "The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver." Hum Mol Genet **18**(21): 4081-4088.

Bell, G. I., K. S. Xiang, et al. (1991). "Gene for non-insulin-dependent diabetes mellitus (maturity-onset diabetes of the young subtype) is linked to DNA polymorphism on human chromosome 20q." Proc Natl Acad Sci U S A **88**(4): 1484-1488.

Blacher, J., R. Asmar, et al. (1999). "Aortic pulse wave velocity as a marker of cardiovascular risk in hypertensive patients." Hypertension **33**(5): 1111-1117.

Bouatia-Naji, N., A. Bonnefond, et al. (2009). "A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk." Nat Genet **41**(1): 89-94.

Bougneres, P. (2003). "Genetics of common obesity and type 2 diabetes: please forget diseases and study pathogenic traits." Diabetes Metab **29**(3): 197-199.

Boule, N. G., E. Haddad, et al. (2001). "Effects of exercise on glycemic control and body mass in type 2 diabetes mellitus: a meta-analysis of controlled clinical trials." JAMA **286**(10): 1218-1227.

Broman, K. W., J. C. Murray, et al. (1998). "Comprehensive human genetic maps: individual and sex-specific variation in recombination." Am J Hum Genet **63**(3): 861-869.

Brull, D. J., L. J. Murray, et al. (2001). "Effect of a COL1A1 Sp1 binding site polymorphism on arterial pulse wave velocity: an index of compliance." Hypertension **38**(3): 444-448.

Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association." Lancet **361**(9357): 598-604.

Carey, G. and J. Williamson (1991). "Linkage analysis of quantitative traits: increased power by using selected samples." Am J Hum Genet **49**(4): 786-796.

Carter, C. O. (1969). "Genetics of common disorders." Br Med Bull **25**(1): 52-57.

Chagnon, Y. C., W. J. Chen, et al. (1997). "Linkage and association studies between the melanocortin receptors 4 and 5 genes and obesity-related phenotypes in the Quebec Family Study." Mol Med **3**(10): 663-673.

Champagne, C. M. (2009). "The usefulness of a Mediterranean-based diet in individuals with type 2 diabetes." Curr Diab Rep **9**(5): 389-395.

Chen, C. H., E. Nevo, et al. (1997). "Estimation of central aortic pressure waveform by mathematical transformation of radial tonometry pressure. Validation of generalized transfer function." Circulation **95**(7): 1827-1836.

Chen, W. M. and G. R. Abecasis (2007). "Family-based association tests for genomewide association scans." Am J Hum Genet **81**(5): 913-926.

Chen, W. M., M. R. Erdos, et al. (2008). "Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels." J Clin Invest **118**(7): 2620-2628.

Cho, Y. S., M. J. Go, et al. (2009). "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits." Nat Genet **41**(5): 527-534.

Chung, W. K., L. Deng, et al. (2009). "Polymorphism in the angiotensin II type 1 receptor (AGTR1) is associated with age at diagnosis in pulmonary arterial hypertension." J Heart Lung Transplant **28**(4): 373-379.

Consoli, A., R. Gomis, et al. (2004). "Initiating oral glucose-lowering therapy with metformin in type 2 diabetic patients: an evidence-based strategy to reduce the burden of late-developing diabetes complications." Diabetes Metab **30**(6): 509-516.

Consortium, H. (2003). "The International HapMap Project." Nature **426**(6968): 789-796.
Craig, M. E., A. Hattersley, et al. (2006). "ISPAD Clinical Practice Consensus Guidelines 2006-2007. Definition, epidemiology and classification." Pediatr Diabetes **7**(6): 343-351.
Crilly, M., C. Coch, et al. (2007). "Repeatability of central aortic blood pressures measured non-invasively using radial artery applanation tonometry and peripheral pulse wave analysis." Blood Press **16**(4): 262-269.
da Silva Xavier, G., M. K. Loder, et al. (2009). "TCF7L2 regulates late events in insulin secretion from pancreatic islet beta-cells." Diabetes **58**(4): 894-905.
Devlin, B., S. A. Bacanu, et al. (2004). "Genomic Control to the extreme." Nat Genet **36**(11): 1129-1130; author reply 1131.
Devlin, B. and K. Roeder (1999). "Genomic control for association studies." Biometrics **55**(4): 997-1004.
Dib, C., S. Faure, et al. (1996). "A comprehensive genetic map of the human genome based on 5,264 microsatellites." Nature **380**(6570): 152-154.
Dupuis, J., C. Langenberg, et al. (2010). "New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk." Nat Genet **42**(2): 105-116.
Eckert, K. A. and S. E. Hile (2009). "Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome." Mol Carcinog **48**(4): 379-388.
Eden, E. R., R. P. Naoumova, et al. (2001). "Use of homozygosity mapping to identify a region on chromosome 1 bearing a defective gene that causes autosomal recessive homozygous hypercholesterolemia in two unrelated families." Am J Hum Genet **68**(3): 653-660.
Elgar, G. and T. Vavouri (2008). "Tuning in to the signals: noncoding sequence conservation in vertebrate genomes." Trends Genet **24**(7): 344-352.
Esposito, K., M. Ciotola, et al. (2007). "Mediterranean diet and the metabolic syndrome." Mol Nutr Food Res **51**(10): 1268-1274.
Fatemi, M., M. M. Pao, et al. (2005). "Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level." Nucleic Acids Res **33**(20): e176.
Fawcett, K. A. and I. Barroso (2010). "The genetics of obesity: FTO leads the way." Trends Genet **26**(6): 266-274.
Franklin, C. S., Y. S. Aulchenko, et al. (2010). "The TCF7L2 diabetes risk variant is associated with HbA(C) levels: a genome-wide association meta-analysis." Ann Hum Genet **74**(6): 471-478.
Gay, J., S. Myers, et al. (2007). "Estimating meiotic gene conversion rates from population genetic data." Genetics **177**(2): 881-894.
Gloyn, A. L. (2003). "Glucokinase (GCK) mutations in hyper- and hypoglycemia: maturity-onset diabetes of the young, permanent

neonatal diabetes, and hyperinsulinemia of infancy." Hum Mutat **22**(5): 353-362.
Grant, S. F. A., G. Thorleifsson, et al. (2006). "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes." Nature Genetics **38**(3): 320-323.
Hakim, O., S. John, et al. (2009). "Glucocorticoid receptor activation of the Ciz1-Lcn2 locus by long range interactions." J Biol Chem **284**(10): 6048-6052.
Hakim, S. A., G. N. Vyas, et al. (1961). "Eleven cases of "Bombay" phenotype in six families: suppression of ABO antigen demonstrated in two families." Transfusion **1**: 218-222.
Haldane, J. B. (2004). "The rate of spontaneous mutation of a human gene. 1935." J Genet **83**(3): 235-244.
Hammoud, S. S., D. A. Nix, et al. (2009). "Distinctive chromatin in human sperm packages genes for embryo development." Nature **460**(7254): 473-478.
Handler, J. (2009). "The importance of accurate blood pressure measurement." Perm J **13**(3): 51-54.
Hanon, O., V. Luong, et al. (2001). "Aging, carotid artery distensibility, and the Ser422Gly elastin gene polymorphism in humans." Hypertension **38**(5): 1185-1189.
Hardison, R., D. Krane, et al. (1991). "Sequence and comparative analysis of the rabbit alpha-like globin gene cluster reveals a rapid mode of evolution in a G + C-rich region of mammalian genomes." J Mol Biol **222**(2): 233-249.
Hardison, R. C. (2000). "Conserved noncoding sequences are reliable guides to regulatory elements." Trends Genet **16**(9): 369-372.
He, J. and P. K. Whelton (1997). "Epidemiology and prevention of hypertension." Med Clin North Am **81**(5): 1077-1097.
Herder, C., W. Rathmann, et al. (2008). "Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 Genes Confer Risk of Type 2 Diabetes Independently of BMI in the German KORA Studies." Horm Metab Res.
Heutink, P. and B. A. Oostra (2002). "Gene finding in genetically isolated populations." Hum Mol Genet **11**(20): 2507-2515.
Hirasawa, R. and R. Feil (2010). "Genomic imprinting and human disease." Essays Biochem **48**(1): 187-200.
Holstein, A., M. Hahn, et al. (2011). "TCF7L2 and therapeutic response to sulfonylureas in patients with type 2 diabetes." BMC Med Genet **12**: 30.
Hutton, J. C. and R. M. O'Brien (2009). "Glucose-6-phosphatase catalytic subunit gene family." J Biol Chem **284**(43): 29241-29245.
Imperatore, G., R. L. Hanson, et al. (1998). "Sib-pair linkage analysis for susceptibility genes for microvascular complications among Pima Indians with type 2 diabetes. Pima Diabetes Genes Group." Diabetes **47**(5): 821-830.

Johansson, A., V. Vavruch-Nilsson, et al. (2005). "Linkage disequilibrium between microsatellite markers in the Swedish Sami relative to a worldwide selection of populations." Hum Genet **116**(1-2): 105-113.
Jones, C. T., I. McIntosh, et al. (1992). "Three novel mutations in the cystic fibrosis gene detected by chemical cleavage: analysis of variant splicing and a nonsense mutation." Hum Mol Genet **1**(1): 11-17.
Kearney, P. M., M. Whelton, et al. (2005). "Global burden of hypertension: analysis of worldwide data." Lancet **365**(9455): 217-223.
Kerem, B., J. M. Rommens, et al. (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-1080.
Kimber, C. H., A. S. Doney, et al. (2007). "TCF7L2 in the Go-DARTS study: evidence for a gene dose effect on both diabetes susceptibility and control of glucose levels." Diabetologia **50**(6): 1186-1191.
Kirichenko, A. V., N. M. Belonogova, et al. (2009). "PedStr software for cutting large pedigrees for haplotyping, IBD computation and multipoint linkage analysis." Ann Hum Genet **73**(Pt 5): 527-531.
Kong, A., D. F. Gudbjartsson, et al. (2002). "A high-resolution recombination map of the human genome." Nat Genet **31**(3): 241-247.
Kong, A., V. Steinthorsdottir, et al. (2009). "Parental origin of sequence variants associated with complex diseases." Nature **462**(7275): 868-874.
Konoshita, T., N. Kato, et al. (2009). "Genetic variant of the Renin-Angiotensin system and diabetes influences blood pressure response to Angiotensin receptor blockers." Diabetes Care **32**(8): 1485-1490.
Kraft, P. (2008). "Curses--winner's and otherwise--in genetic epidemiology." Epidemiology **19**(5): 649-651; discussion 657-648.
Krawetz, S. A. (2005). "Paternal contribution: new insights and future challenges." Nat Rev Genet **6**(8): 633-642.
Krentz, A. J. and C. J. Bailey (2005). "Oral antidiabetic agents: current role in type 2 diabetes mellitus." Drugs **65**(3): 385-411.
Lander, E. and L. Kruglyak (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nat Genet **11**(3): 241-247.
Lango, H., C. N. Palmer, et al. (2008). "Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk." Diabetes.
Latini, V., G. Sole, et al. (2004). "Genetic isolates in Corsica (France): linkage disequilibrium extension analysis on the Xq13 region." Eur J Hum Genet **12**(8): 613-619.
Lehman, D. M., K. J. Hunt, et al. (2007). "Haplotypes of transcription factor 7-like 2 (TCF7L2) gene and its upstream region are associated with type 2 diabetes and age of onset in Mexican Americans." Diabetes **56**(2): 389-393.
Levy, D., G. B. Ehret, et al. (2009). "Genome-wide association study of blood pressure and hypertension." Nat Genet.

Levy, D., M. G. Larson, et al. (2007). "Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness." BMC Med Genet **8 Suppl 1**: S3.

Lynch, M. (2010). "Rate, molecular spectrum, and consequences of human mutation." Proc Natl Acad Sci U S A **107**(3): 961-968.

Macgregor, S., B. K. Cornes, et al. (2006). "Bias, precision and heritability of self-reported and clinically measured height in Australian twins." Hum Genet **120**(4): 571-580.

Maher, B. (2008). "Personal genomes: The case of the missing heritability." Nature **456**(7218): 18-21.

Manley, S., W. G. John, et al. (2004). "Introduction of IFCC reference method for calibration of HbA: implications for clinical care." Diabet Med **21**(7): 673-676.

Marchini, J., L. R. Cardon, et al. (2004). "The effects of human population structure on large genetic association studies." Nat Genet **36**(5): 512-517.

Matthews, D. R., J. P. Hosker, et al. (1985). "Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man." Diabetologia **28**(7): 412-419.

Mayfield, J. A. and R. D. White (2004). "Insulin therapy for type 2 diabetes: rescue, augmentation, and replacement of beta-cell function." Am Fam Physician **70**(3): 489-500.

McKnight, J. A., A. D. Morris, et al. (2008). "Implementing a national quality assurance system for diabetes care: the Scottish Diabetes Survey 2001-2006." Diabet Med **25**(6): 743-746.

McQuillan, R., A. L. Leutenegger, et al. (2008). "Runs of homozygosity in European populations." Am J Hum Genet **83**(3): 359-372.

Medley, T. L., T. J. Cole, et al. (2002). "Fibrillin-1 genotype is associated with aortic stiffness and disease severity in patients with coronary artery disease." Circulation **105**(7): 810-815.

Meigs, J. B., A. K. Manning, et al. (2007). "Genome-wide association with diabetes-related traits in the Framingham Heart Study." BMC Med Genet **8 Suppl 1**: S16.

Murray, J. C., K. H. Buetow, et al. (1994). "A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC)." Science **265**(5181): 2049-2054.

Newton-Cheh, C., T. Johnson, et al. (2009). "Genome-wide association study identifies eight loci associated with blood pressure." Nat Genet.

Ng, M. C., K. S. Park, et al. (2008). "Implication of Genetic Variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2 and FTO in Type 2 Diabetes and Obesity in 6719 Asians." Diabetes.

O'Rourke, M. (1990). "Arterial stiffness, systolic blood pressure, and logical treatment of arterial hypertension." Hypertension **15**(4): 339-347.

O'Rourke, M. F., A. Pauca, et al. (2001). "Pulse wave analysis." Br J Clin Pharmacol **51**(6): 507-522.

Org, E., S. Eyheramendy, et al. (2009). "Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations." Hum Mol Genet **18**(12): 2288-2296.

Papaioannou, T. G., A. D. Protogerou, et al. (2009). "Non-invasive methods and techniques for central blood pressure estimation: procedures, validation, reproducibility and limitations." Curr Pharm Des **15**(3): 245-253.

Pardo, L. M., I. MacKay, et al. (2005). "The effect of genetic drift in a young genetically isolated population." Ann Hum Genet **69**(Pt 3): 288-295.

Pare, G., D. I. Chasman, et al. (2008). "Novel association of HK1 with glycated hemoglobin in a non-diabetic population: a genome-wide evaluation of 14,618 participants in the Women's Genome Health Study." PLoS Genet **4**(12): e1000312.

Pattaro, C., F. Marroni, et al. (2007). "The genetic study of three population microisolates in South Tyrol (MICROS): study design and epidemiological perspectives." BMC Med Genet **8**: 29.

Pauca, A. L., M. F. O'Rourke, et al. (2001). "Prospective evaluation of a method for estimating ascending aortic pressure from the radial artery pressure waveform." Hypertension **38**(4): 932-937.

Pearson, E. R., L. A. Donnelly, et al. (2007). "Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study." Diabetes **56**(8): 2178-2182.

Peltonen, L., A. Palotie, et al. (2000). "Use of population isolates for mapping complex traits." Nat Rev Genet **1**(3): 182-190.

Perez-Lopez, F. R., P. Chedraui, et al. (2009). "Effects of the Mediterranean diet on longevity and age-related morbid conditions." Maturitas **64**(2): 67-79.

Pilia, G., W. M. Chen, et al. (2006). "Heritability of cardiovascular and personality traits in 6,148 Sardinians." PLoS Genet **2**(8): e132.

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-909.

Proietti, A. R., V. del Balzo, et al. (2009). "[Mediterranean diet and prevention of non-communicable diseases: scientific evidences]." Ann Ig **21**(3): 197-210.

Prokopenko, I., C. Langenberg, et al. (2009). "Variants in MTNR1B influence fasting glucose levels." Nat Genet **41**(1): 77-81.

Purcell, S., S. S. Cherny, et al. (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits." Bioinformatics **19**(1): 149-150.

Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.

Ramracheya, R. D., D. S. Muller, et al. (2008). "Function and expression of melatonin receptors on human pancreatic islets." J Pineal Res **44**(3): 273-279.

Rassoulzadegan, M., V. Grandjean, et al. (2006). "RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse." Nature **441**(7092): 469-474.

Rodrigues, R., C. S. Gabetta, et al. (2008). "Cystic fibrosis and neonatal screening." Cad Saude Publica **24 Suppl 4**: s475-484.

Roth, C. L., A. Hinney, et al. (2008). "TCF7L2 Polymorphism rs7903146 and Predisposition for Type 2 Diabetes Mellitus in Obese Children." Horm Metab Res.

Rowe, S. M., S. Miller, et al. (2005). "Cystic fibrosis." N Engl J Med **352**(19): 1992-2001.

Sanghera, D. K., L. Ortega, et al. (2008). "Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk." BMC Med Genet **9**(1): 59.

Sheffield, V. C., E. M. Stone, et al. (1998). "Use of isolated inbred human populations for identification of disease genes." Trends Genet **14**(10): 391-396.

Shifman, S. and A. Darvasi (2001). "The value of isolated populations." Nat Genet **28**(4): 309-310.

Silbernagel, G., W. Renner, et al. (2011). "Association of TCF7L2 SNPs with Age of Onset of Type 2 Diabetes and Proinsulin/Insulin Ratio but not with Glucagon Like Peptide 1." Diabetes Metab Res Rev.

Sladek, R., G. Rocheleau, et al. (2007). "A genome-wide association study identifies novel risk loci for type 2 diabetes." Nature **445**(7130): 881-885.

Soranzo, N., S. Sanna, et al. (2010). "Common variants at ten genomic loci influence hemoglobin A1C levels via glycemic and non-glycemic pathways." Diabetes.

Southam, L., K. Panoutsopoulou, et al. (2011). "The effect of genome-wide association scan quality control on imputation outcome for common variants." Eur J Hum Genet **19**(5): 610-614.

Starr, I., C. Ambrosi, et al. (1973). "Diagnosis of aortic stenosis from the carotid pulse and its derivative." Br Heart J **35**(10): 1062-1065.

Stumpf, I., E. Muhlbauer, et al. (2008). "Involvement of the cGMP pathway in mediating the insulin-inhibitory effect of melatonin in pancreatic beta-cells." J Pineal Res **45**(3): 318-327.

Tian, D., Q. Wang, et al. (2008). "Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes." Nature **455**(7209): 105-108.

Tobin, M. D., N. A. Sheehan, et al. (2005). "Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure." Stat Med **24**(19): 2911-2935.

Tohidi, M., M. Hatami, et al. (2010). "Lipid measures for prediction of incident cardiovascular disease in diabetic and non-diabetic adults: results of the 8.6 years follow-up of a population based cohort study." Lipids Health Dis **9**: 6.

Urquiaga, I., G. Echeverria, et al. (2008). "Mediterranean food and diets, global resource for the control of metabolic syndrome and chronic diseases." World Rev Nutr Diet **98**: 150-173.

Vitart, V., Z. Biloglav, et al. (2006). "3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia." Eur J Hum Genet **14**(4): 478-487.

Vitart, V., I. Rudan, et al. (2008). "SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout." Nat Genet **40**(4): 437-442.

Voight, B. F., L. J. Scott, et al. (2010). "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis." Nat Genet **42**(7): 579-589.

Wagner, K. D., N. Wagner, et al. (2008). "RNA induction and inheritance of epigenetic cardiac hypertrophy in the mouse." Dev Cell **14**(6): 962-969.

Wang, Y., J. R. O'Connell, et al. (2009). "From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene." Proc Natl Acad Sci U S A **106**(1): 226-231.

Weale, M. E. (2010). "Quality control for genome-wide association studies." Methods Mol Biol **628**: 341-372.

Weijnen, C. F., S. S. Rich, et al. (2002). "Risk of diabetes in siblings of index cases with Type 2 diabetes: implications for genetic studies." Diabet Med **19**(1): 41-50.

WHO (2006). "Definition and Diagnosis or Diabetes Mellitus and Intermediate Hyperglycemia: Report of a WHO/IDF consultation."

WHO (2008). The global burden of disease: 2004 update.

Wild, S., G. Roglic, et al. (2004). "Global prevalence of diabetes: estimates for the year 2000 and projections for 2030." Diabetes Care **27**(5): 1047-1053.

Williams, J. T. and J. Blangero (1999). "Power of variance component linkage analysis to detect quantitative trait loci." Ann Hum Genet **63**(Pt 6): 545-563.

Wollstein, A., A. Herrmann, et al. (2007). "Efficacy assessment of SNP sets for genome-wide disease association studies." Nucleic Acids Res **35**(17): e113.

Woodward, M. and H. Tunstall-Pedoe (1995). "Alcohol consumption, diet, coronary risk factors, and prevalent coronary heart disease in men and women in the Scottish heart health study." J Epidemiol Community Health **49**(4): 354-362.

Wright, A. F., A. D. Carothers, et al. (1999). "Population choice in mapping genes for complex diseases." Nat Genet **23**(4): 397-404.

WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

Yasmin and K. M. O'Shaughnessy (2008). "Genetics of arterial structure and function: towards new biomarkers for aortic stiffness?" Clin Sci (Lond) **114**(11): 661-677.

Zeggini, E., L. J. Scott, et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nat Genet **40**(5): 638-645.

8 Appendix I. HbA_{1C} Publication

The TCF7L2 diabetes risk variant is associated with HbA_1(C) levels: a genome-wide association meta-analysis.

Franklin CS, Aulchenko YS, Huffman JE, Vitart V, Hayward C, Polašek O, Knott S, Zgaga L, Zemunik T, Rudan I, Campbell H, Wright AF, Wild SH, Wilson JF.

Ann Hum Genet. 2010 Nov;**74**(6):471-8